Faulty Neural Networks

1st Shiuan-Wen Chen
Dept. of Electrical and Computer Engineering (ECE)
University of Toronto
Toronto, Canada
shiuanwen.chen@mail.utoronto.ca

2nd Brendan Duke
Dept. of ECE
University of Toronto
Toronto, Canada
brendan.duke@utoronto.ca

3rd Parham Aarabi
Dept. of ECE
University of Toronto
Toronto, Canada
p@arh.am

***Abstract*—This study aims to investigate the response of the nervous system to injury through experiments using a neural network model trained with the MNIST dataset [1]. Multiple experiments are performed to examine the relationship between neural network damage and accuracy. How the damaged network can restore its functionality or accuracy with the aid of another neural network is also investigated. By analyzing these results, a better understanding of the nervous system's ability to respond to injury and adapt to changes in neural networks can be gained.**

***Index Terms*—neural network, neural injury, artificial intelligence**

I. Introduction

In recent years, neural networks have gained widespread adoption as tools for modeling and understanding the human brain. However, the response of the neural system to injury remains a challenging problem. To address this issue, researchers have conducted a range of experiments that involve damage trained neural network. Meyes et al. performed experiments to examine the robustness versus structural damage in the neural network.[2] Cun et al. discovered that removing unimportant weight improves the learning speed and slightly in the accuracy. [3] Such experiments offer valuable insights into the impact of the neural injury on neuron function, and may ultimately lead to the development of novel approaches for treating neural injuries.

This study investigates the impact of simulated neural injury on a trained neural network with MNIST dataset using two distinct experimental approaches. The first approach selectively damages specific neurons, and the second approach selectively damages the outputs of specific layer of neurons. The network's performance is evaluated after simulated neural injury, providing insight into how the brain responds to injury and adapts to changes in neural networks. The findings contribute to the growing body of literature exploring the impact of neural injury on neural networks and offer potential new avenues for treating neural injuries.

Lastly, the neuron damage experiments were reversed to investigate whether a neural network could repair a damaged network. Such an approach may provide insights into the self-repair mechanisms of the brain and inspire the development of new treatment strategies for neural injuries.

II. Experiments

A. Noisy neuron

To replicate the characteristics of an injured nervous system, the experiment employs a fully connected two-layer neural network comprising an input layer and an output layer. The network is evaluated based on two critical parameters - noise probability and noise level - in terms of accuracy. Additionally, the study introduces two types of damage, namely neuron damage and output damage. The former impacts all outputs from the affected neuron, while the latter only affects certain outputs from the neural layer.

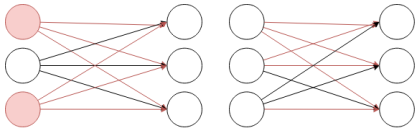

Fig. 1. Neuron damage (LHS) and Output damage (RHS). The red neuron and output are those affected

In Figure 1, both networks exhibit a 33% noise probability in their first layer. However, the network on the left-hand side has incurred neuron damage, resulting in the 1st and 3rd neurons being affected, and consequently, all their outputs being replaced with the same noise. On the other hand, the network on the right-hand side has experienced output damage, where 33% of its outputs in the first layer have been impacted. In this example, as there are a total of 9 outputs from the first layer, 6 outputs have been affected by the damage, which may or may not originate from the same neuron and the noise may or may not be the same. Despite some neurons being impacted, they are still capable of transmitting a portion of the correct signals instead of producing all noise.

$$Noise(i) = X(i) \times noise\_level \tag{1}$$

The relationship between the noise level and the noise is depicted in Eq. 1, where $X \sim \mathcal{N}(0,\, 1)$ and i is either the index of a neuron or index of output depending on the experiment performed. Please notes that the affected outputs or neurons has their value replaced, not added, by the noise.

B. Neural repair

A comprehensive set of experiments is conducted to explore the effects of varying combinations of noise probability and

level on neuron damage. Additionally, a restorative experiment is conducted, wherein noise is added to all the outputs of the first layer, while leaving the underlying structure of the neural network unaltered. A three-layer neural network model, which possesses complete knowledge of both the input and label, has partial access to alter the output per neuron as in neuron damage. This access is quantified by the term "neuron access density" which is similar to the concept of noise probability.

III. Results

A. Noisy neuron

Figure 2 depicts the outcome of the two-layer neuron damage experiment, which examines the impact of different noise levels and probabilities on accuracy. As the noise probabilities increase, more neurons are either injured or affected by the noise, leading to a decrease in accuracy. Moreover, the accuracy decreases at a faster rate as the noise level increases.

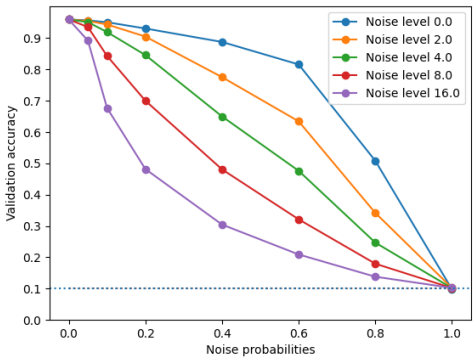

Fig. 2. Noise Probability vs. Accuracy with different noise levels on an entire neuron of a single layer in a two-layer neural network model

Figure 3 depicts similar results to those presented in Figure 2. However, in the former, the accuracy drops at a much faster rate. One interesting observation is that noise level 0.0 decreases more rapidly than noise levels 2.0 and 4.0. This phenomenon is possibly due to the weight in the next layer being unable to propagate since the weight is multiplied by 0, and the bias being insufficient for the model to arrive at the correct answer.

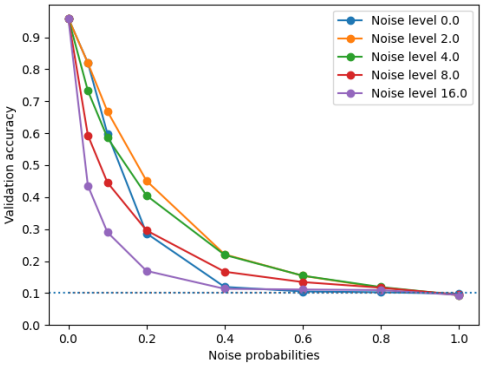

Fig. 3. Noise Probability vs. Accuracy with different levels of noise level on a percentage of outputs of a single layer in 2 layers neural network model

B. Neural repair

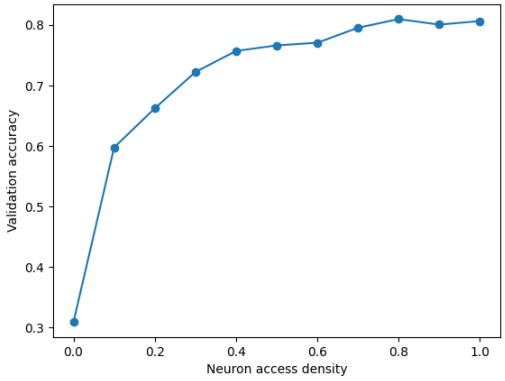

Fig. 4. Noise access density vs. Accuracy with 1 epoch on two-layer neural network model

The findings presented in Figure 4 demonstrate the efficacy of neuron repair in restoring accuracy to neural networks. Accuracy improves as the neuron access density increases, highlighting the viability of repairing damaged or inaccessible neurons. The restoration neural network was able to achieve significant accuracy improvements even after only one epoch, indicating the efficiency and effectiveness of the repair mechanism, and thus showing the repair mechanism has the potential to be dynamic in the human nervous system.

IV. Conclusion

The study aimed to investigate the correlation between the probability and level of noise and their respective effects on accuracy. The results revealed a significant decline in accuracy with an increase in noise probability, and a steeper decline was observed with an elevated noise level. The findings also indicated that the output damage had a more pronounced effect on accuracy than neuron damage, despite having identical levels and probabilities of noise. The neural repair is able to successfully restore functionality in the damaged model. The accuracy of the damaged model increases as the restoration model has more access to the damaged model. The implications of these observations provide valuable insights into the intricate mechanisms underlying the impact of injuries on the nervous system.

References

[1] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[2] R. Meyes, M. Lu, C. W. de Puiseau, and T. Meisen, *Ablation studies in artificial neural networks*, 2019. DOI: 10.48550/ARXIV.1901.08644. [Online]. Available: https://arxiv.org/abs/1901.08644.

[3] Y. Lecun, J. Denker, and S. Solla, "Optimal brain damage," vol. 2, Jan. 1989, pp. 598–605.