VISUALIZING NEURAL NETS

A PREPRINT

Daniel Fishbein
Department of Computer Science
Rochester Institute of Technology
Rochester, NY 14623
df3622@g.rit.edu

Jivitesh Debata
Department of Computer Science
Rochester Institute of Technology
Rochester, NY 14623
jd9039@g.rit.edu

May 9, 2023

ABSTRACT

Artificial Intelligence (AI) models, specifically neural networks (NN), have shown promising potential across various fields, but their failure points in mission-critical applications can lead to catastrophic consequences. This paper explores these limitations through the lens of visualization and boundary conditions, diving into the inner workings of NNs and highlighting the 'black box' problem. We critically discuss existing visualization techniques and their insufficiency to reveal potential failure points. Our study focuses on visualizing and analyzing the minima reached by a NN during its training, providing insights into the network's learned representation and its limitations. We propose a method that engages the network with well-constrained tasks, allowing us to map the decision space of the NN and visualize the minima. Our observations reveal interesting phenomena, such as overlapping decision boundaries and dormant neurons, which can cause vulnerabilities in mission-critical applications. This research aims to contribute to the development of robust, reliable, and transparent AI systems, and provide valuable insights for their responsible deployment.

Keywords Artificial Intelligence (AI) · Neural Networks (NN) · Black box problem · Minima · Decision space · Decision boundaries · Dormant neurons

1 Introduction

Artificial Intelligence (AI) is driving a paradigm shift across many fields, promising unprecedented efficiencies and capabilities. However, this potential is interwoven with numerous challenges. One particular challenge that is often overlooked are failure points in AI models. This paper aims to delve into these issues with the focus on how AI models, specifically neural networks (NN), function and where they falter. The understanding of failure points is especially important when they can cause catastrophic failure to an entire system. We dub these points of potential failure as "mission-critical applications".

The landscape of AI research has seen significant advancements in recent years. Issues relating to the open-set and closed-set problems have been thoroughly explored. Open-set problems, where an AI model has to classify inputs into known classes or declare them as unknown, were discussed by Scheirer et al. in their seminal work [1]. On the other hand, closed-set problems, where all classes are known at the time of training, have also been extensively studied[2].

Neural networks, which form the backbone of most AI systems, have also been the subject of substantial research. The application of Spline theory in neural networks, as proposed by Unser [3], has provided critical insights into their inner workings. Similarly, the representation of neural networks using full gradient methods, a topic explored by Balduzzi et al. [4], has opened up new avenues for understanding these complex models.

While these methods offer better insights into how neural networks learn and express patterns, they do not entirely address the inherent 'black box' nature of said NNs. Efforts towards visualizing what a NN has learned, such as those by Zeiler and Fergus [5], and data-driven visualization of NNs, as seen in the work by Liu et al. [6], have made substantial progress towards this end. Yet, these approaches primarily focus on better expressing what a NN has learned, without deeply probing its limitations or potential failure points.

With the prominent success of AI models many applications are adopting them in contexts that will interact with humans. These NN's are even being applied to mission-critical applications with their designers knowing that it is unknown when the NN will fail. Given the raised states of these situations, understanding NN limitations is paramount. This paper seeks to examine the weaknesses of AI models in the context of their increased adoption in mission-critical applications. Our research will explore the predictability of NN failure through the lens of visualization and the use of boundary conditions. The results of this study show comprehensive shortcomings of NN's and imitate areas of improvement in these AI systems.

Our goal is to contribute to the development of more robust, reliable, and transparent AI systems. This research will provide valuable insights to researchers, practitioners, and policymakers, facilitating the responsible and effective deployment of AI technologies.

2 Related Work

Understanding and interpreting the inner workings of neural networks is a critical aspect of AI research. A prominent part of this field involves visualizing the learning process of these models, with notable contributions from researchers such as Zeiler, Fergus, and Liu.

Zeiler and Fergus [5] developed a visualization technique that helped to understand what features a convolutional network had learned during training. By synthesizing inputs that maximally activate specific neurons, they were able to understand the different layers of a network. Their work effectively enabled the interpretation of individual neuron functions within the network. This work also greatly contributed to the understanding of how different layers contribute to the final output.

Similarly, Liu et al. [6] proposed a data-driven approach to visualize neural networks. Their work involved using an optimization-based technique to reverse-engineer learned representations. This process allowed for a detailed analysis of the activation patterns in neural networks. This work provided further insight into the model's decision-making process.

While both these approaches significantly contributed to the field, they primarily focused on expressing the patterns that the neural networks learned. The visualizations often appeared in the form of what can be interpreted as 'heat maps' - representations of the most travelled paths by an input impulse through the network. These 'heat maps' effectively highlight the neurons or areas of the neural network that are most activated by specific inputs, providing a sense of the patterns or features that the network has learned to recognize.

However, these visualizations tend to focus on the positive aspects of a network's learning process. Put another way, "What has the network learned and how it has learned it". These positively focused studies neglected failure states in NN's and will not help us identify when a given NN will fail. Our work aims to build on these efforts and explore the limitations and potential failure points of these models.

3 Problem description

The primary problem we aim to address in this study is the visualization of neural networks. At first glance, this task seems quite straightforward. However, when considering the high dimensionality of neural networks, especially deep networks, this becomes a complex challenge.

Typical visualization techniques focus on the weights and biases of the neural network, which although insightful, do not provide a complete understanding of the model's learning process or its potential limitations. The challenge lies in not just visualizing the state of the neural network but also in deciphering the implications of that state. In particular, the current minima that the network's optimization process has reached.

In light of this, we redefine our problem: instead of attempting to visualize the entire neural network, we aim to visualize and analyze the current minima that the neural network has arrived at during its training process. This minima, a point in the high-dimensional space of the network's parameters, can be seen as a representation of what the network has learned.

By visualizing the minima in this manner, we aim to explore its properties. We encourage the reader to imagine standing on a hilltop and surveying the landscape. Each hill is a region of high dimensional space that has been labeled with a corresponding class by the NN. This analogy allows us to traverse the terrain created by the NN and by extension gives us a deeper insight into the regions that the NN has defined i.e, "learned".

Therefore, the core problem our study tackles is the visualization and analysis of the minima reached by a neural network during its training. This approach provides a more intuitive understanding of the network's learned representation, facilitating insights into its limitations and potential failure points.

4 Proposed solution

Our proposed solution to the problem of visualizing and understanding the current minima of a neural network involves a systematic and well-structured experimental setup. We aim to engage the network with a variety of mathematically defined tasks that are well-constrained, and which the network can learn to solve.

The tasks are designed in a manner that allows for a clear understanding of what the network has learned. This in turn provides insights into the minima it achieved. By examining the output of individual neurons, we can investigate the internal processes of the network and its learned outputs.

This process allows us to map the decision space of the neural network to the corresponding inputs. By graphing these input-output pairs we generate a visual representation of the minima that has been achieved during training. The analysis of this minima and the corresponding neuron activation's presents a unique approach to understanding the network's learned minima and identify potential limitations that could cause points of failure.

5 Experimental design

For this study, we designed a series of experiments using a small neural network composed of two input neurons, a hidden layer with three neurons, and a final layer with three output neurons.

Our experimental setup involved both mathematical and human-based tests. The human test revolved around the simple task of identifying whether an image is more black, white, or grey, Given two pixel values ranging from 0-255 each. Although this might seem unconventional to human image recognition, it parallels the current process of how neural networks interpret image data. It highlights the difference in inherent cognitive abilities between humans and AI models, providing a unique perspective on the challenges faced by neural networks in understanding and learning from data.

For the mathematical tests, we defined three distinct data sets, S1, S2, and S3. Each set was designed to emphasize certain characteristics and potential pitfalls in the network's learning process while still preserving the question asked of our human annotators.

Data set S1: This set was defined by the following piece wise function:

$$S1(x) = \begin{cases} 'white' & x_1 + x_2 < 255 - n \\ 'Black' & x_1 + x_2 > 255 + n \\ 'Grey' & otherwise \end{cases}$$

Where x_1 and x_2 are the input neurons and n is an arbitrary constant. The design of S1 allowed us to study the network's ability to make binary decisions based on the relative magnitude of the inputs and their sum.

Data set S2: The second set was defined by the following piece wise function:

$$S2(x) = \begin{cases} 'white' & x_1 < x_2 \text{ and } x_1 + x_2 < 255 - n \\ 'Black' & x_1 > x_2 \text{ and } x_1 + x_2 > 255 + n \\ 'Grey' & otherwise \end{cases}$$

Where x_1 and x_2 are the input neurons and n is an arbitrary constant. S2 was designed to explore how the network handled cases with overlapping conditions.

Data set S3: The third set was defined based on the following piece-wise function:

$$S3(x_i) = \begin{cases} 'white' & x_1 + x_2 < 255 - n \\ 'Black' & x_1 + x_2 > 255 + n \\ 'Grey' & x_i = 3 \\ 'Grey' & otherwise \end{cases}$$

Where x_1 and x_2 are the input neurons and n is an arbitrary constant. x_i refers to every third input. Meaning that the value will be "Grey" on every third image. S3 aimed to test the network's ability to handle non-linear and non-continuous data. This set can be loosely thought of as time-series data.

Each of these experiments was tailored to probe different aspects of the neural network's learning and decision-making processes. By assessing the network's performance across these varied and increasingly complex scenarios, we aimed to gain a more nuanced understanding of its limitations and potential failure points.

A network that was run on each data set can be viewed in figures 3, 4, 5, 6.

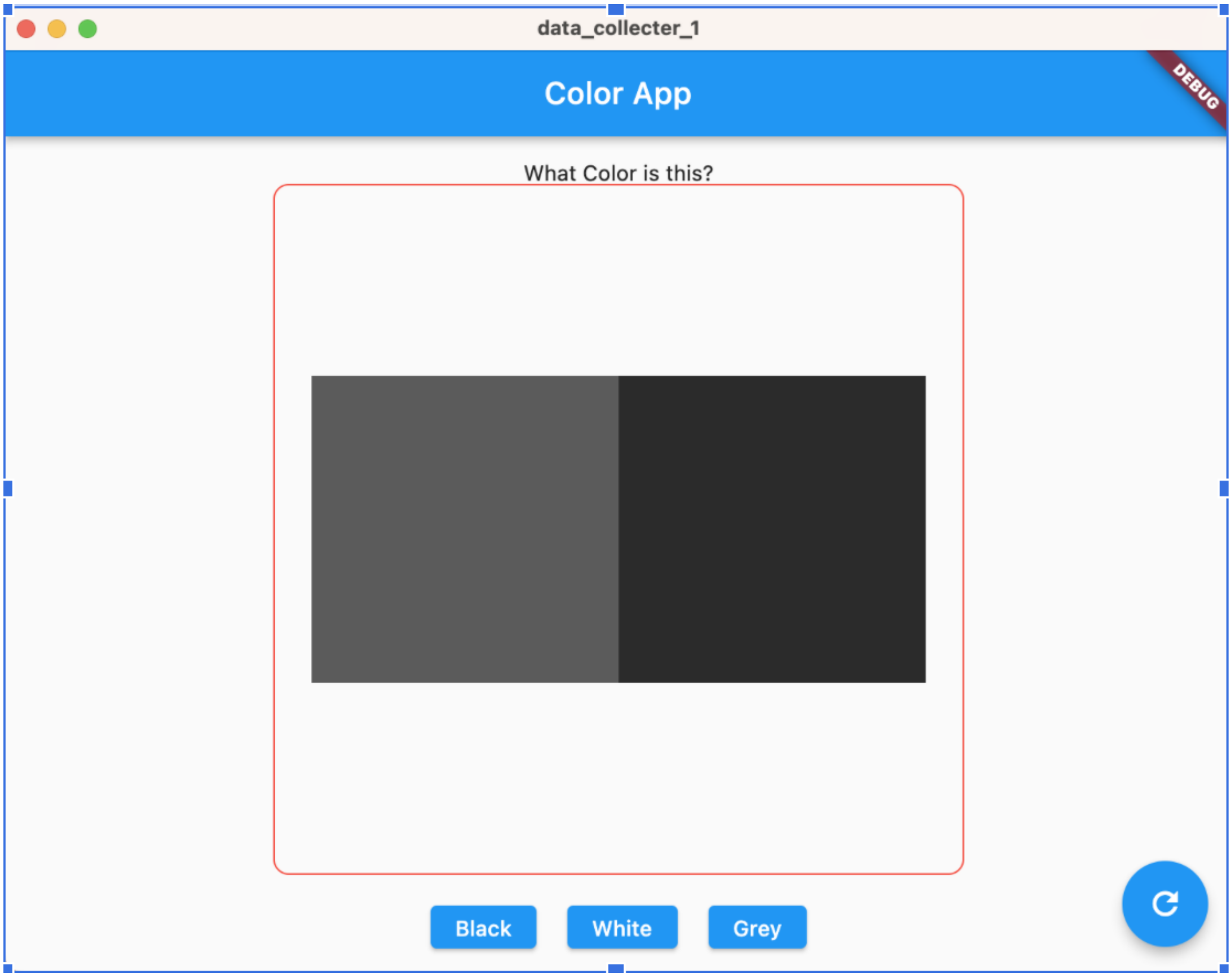

Figure 1: This figure demonstrates how the human annotation data of the Black, White, Grey experiment was gathered. After each answer each Box will be filled with a value between 0 and 255. The use must decide if the image is more black, white or grey.

5.1 The math of a Neural Network:

Visualizing the decision-making process of the model trained on the MNIST data set, which has over 700 dimensions of input, posed a significant challenge. Simple dimensionality reduction techniques are susceptible to a myriad of problems. The least of which is data order as seen in figure 8. These methods can also be disingenuously of the actual shape of the NN boundaries [2].

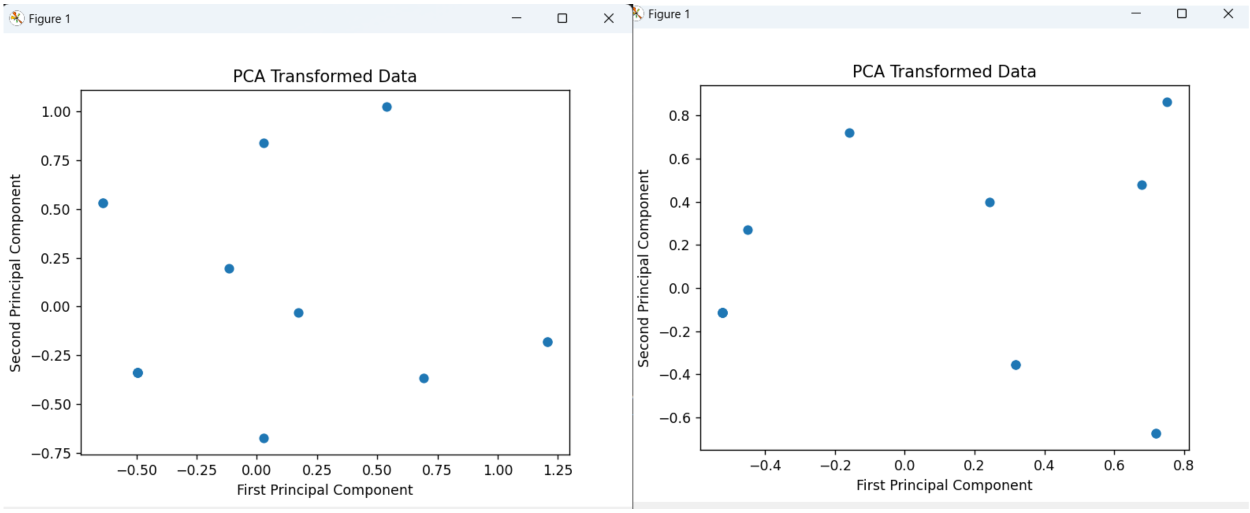

Figure 2: This figure shows a Principal Component Analysis (PCA) of a $1, 1, 1, 1$ hyper cube's vertices. Simply by shuffling the order of the data, these two distinct plots are made.

An approach we attempted to find these boundary conditions was to solve the NN for when 1 class equals another. For the sake the authors and time we only consider a NN with 2 input neurons, 1 hidden layer with 3 neurons and 2 output neurons. The following argument follows:

Given a NN with 2 input neurons, 1 hidden layer with 3 neurons and 2 output neurons, that is fully connected, and uses a sigmoid activation. A diagram of this NN can be seen in figure 9.

When does $z_1 = z_2$?

To elaborate on the notation that will be used see the following definitions. : x_1, x_2 are the input neurons y_1, y_2, y_3 are the hidden layers neurons z_1, z_2 are the output neurons

w denotes a weight
$w_{y_1->z_2}$ denotes the weight from y_1 to z_2

b denotes a bias
b_{y_3} denotes the bias associated with neuron y_3

By tracing the connections we can derive the following equations: The input to a y_i neuron will be denoted as:

$y_i = \sigma(w_{x_1->y_i} * x_1 + w_{x_2->y_i} * x_2 + b_{y_i})$
Where $\sigma(x) = \frac{1}{1+e^{-x}} = \frac{1}{1+\exp[-x]}$

The input to a z_i neuron will be denoted as:
$z_i = \sigma(w_{y_1->z_i} * y_1 + w_{y_2->z_i} * y_2 + w_{y_3->z_i} * y_3 + b_{z_i})$

We now set $z_1 = z_2$. We then expand our expression by substation. This leave us with:

$$
\begin{gathered}
(1 + \exp[-(w_{y_1->z_1} * \frac{1}{1 + \exp[w_{x_1->y_1} * x_1 + w_{x_2->y_1} * x_2 + b_{y_1}]} \\
+ w_{y_2->z_1} * \frac{1}{1 + \exp[w_{x_1->y_2} * x_1 + w_{x_2->y_2} * x_2 + b_{y_2}]} \\
+ w_{y_3->z_1} * \frac{1}{1 + \exp[w_{x_1->y_3} * x_1 + w_{x_2->y_3} * x_2 + b_{y_3}]} \\
+ b_{z_1})])^{-1} \\
= \\
(1 + \exp[-(w_{y_1->z_2} * \frac{1}{1 + \exp[w_{x_1->y_1} * x_1 + w_{x_2->y_1} * x_2 + b_{y_1}]} \\
+ w_{y_2->z_2} * \frac{1}{1 + \exp[w_{x_1->y_2} * x_1 + w_{x_2->y_2} * x_2 + b_{y_2}]} \\
+ w_{y_3->z_2} * \frac{1}{1 + \exp[w_{x_1->y_3} * x_1 + w_{x_2->y_3} * x_2 + b_{y_3}]} \\
+ b_{z_2})])^{-1}
\end{gathered}
\tag{1}
$$

With the use of copious amounts of algebra we can rewrite our equation as:

$$
\begin{gathered}
C_0 + C_1 * \exp[B_1] + C_2 * \exp[B_2] + C_3 * \exp[B_3] + \\
C_4 * \exp[B_2 + B_3] + C_5 * \exp[B_1 + B_3] + C_6 * \exp[B_1 + B_2] + \\
C_7 * \exp[B_1 + B_2 + B_3] = 0
\end{gathered}
\tag{2}
$$

Where $C_0, C_1, ...C_7$ and B_1, B_2, B_3 are defined as follows:

$$
\begin{gathered}
C_0 = w_{y_1->z_1} - w_{y_1->z_2} + w_{y_2->z_1} - w_{y_2->z_2} + w_{y_3->z_1} - w_{y_3->z_2} - b_{z_2} + b_{z_1} \\
C_3 = (w_{y_1->z_1} - w_{y_1->z_2} + w_{y_2->z_1} - w_{y_2->z_2} - b_{z_2} + b_{z_1}) \\
C_2 = (w_{y_1->z_1} - w_{y_1->z_2} + w_{y_3->z_1} - w_{y_3->z_2} - b_{z_2} + b_{z_1}) \\
C_1 = (w_{y_2->z_1} - w_{y_2->z_2} + w_{y_3->z_1} - w_{y_3->z_2} - b_{z_2} + b_{z_1}) \\
C_4 = (w_{y_1->z_1} - w_{y_1->z_2} - b_{z_2} + b_{z_1}) \\
C_5 = (w_{y_2->z_1} - w_{y_2->z_2} - b_{z_2} + b_{z_1}) \\
C_6 = (w_{y_3->z_1} - w_{y_3->z_2} - b_{z_2} + b_{z_1}) \\
C_7 = (-b_{z_2} + b_{z_1})
\end{gathered}
\tag{3}
$$

$$
\begin{gathered}
B_1 = -w_{x_1->y_1} * x_1 + w_{x_2->y_1} * x_2 + b_{y_1} \\
B_2 = -w_{x_1->y_2} * x_1 + w_{x_2->y_2} * x_2 + b_{y_2} \\
B_3 = -w_{x_1->y_3} * x_1 + w_{x_2->y_3} * x_2 + b_{y_3}
\end{gathered}
\tag{4}
$$

Some interesting observations about equation 2 is that all values of C_i are constants, and that B_1, B_2, B_3 is the same equation for a y neuron but without the activation function.

It is also of interesting how ordered and structured equation 2 is.

Future work might explore the use of differential equations as the general solution of:

$$\frac{dy}{dx} = ky \tag{5}$$

is:

$$y = Ce^{kx} = C * \exp[kx] \tag{6}$$

where C and k are constants (Remember that this C different than our C_n)

The gneral solution to this simple first order differential equasion is very close to a series of solutions for each term outlined in equasion 1. By assuming a continious input space, the future work might then use aproxomation techniques for finding the boundery conditions. These aproxomation techniques only need to be good enough since the machine learning problem of classification is discrete.

6 Experimental Results

The conducted experiments produced intriguing results, as detailed in the figures. Figure 1 presents an output shape of a single neuron, illustrating the sigmoid activation function's influence on the given neural network architecture. This architecture, composed of two input neurons, a hidden layer with three neurons, and three output neurons, was utilized throughout the experiments.

Figures 4-7 depict the geometries generated when mapping two input values to their output for each data set (S1 to S4). The distinct structures formed by these mappings highlight the unique characteristics of each data set and offer insight into how the neural network interprets and learns from these different inputs.

From the human annotation data gathered (Figure 6), it was observed that the task of identifying whether an image is more black, white, or grey, given two pixel values ranging from 0-255, was not as straightforward for humans as it might seem for an AI model. This draws attention to the differing cognitive abilities between humans and AI models, providing a unique perspective on the challenges faced by neural networks in understanding and learning from data.

Overall, the experiment results shed light on the nuanced understanding of the limitations and potential failure points of the neural network's learning and decision-making processes, providing a foundation for future investigations.

Each of these figures plots the geometry of input variables to an output variables

7 Discussion

Throughout the course of our experiments, several intriguing observations were made which provide further insight into the learning process of neural networks and their potential limitations.

One significant observation was the overlapping planes of individual neurons. During the operation of the network we noted that the decision boundaries of different neurons often intersected, creating regions of potential ambiguity. These intersection points represent inputs where the network has two distinct outputs. On one side of the boundary is labeled one class and the other side is labeled as a different class. By varying the input such that one travels over a boundary, the NN will change its prediction from one classification to another.

While the comprehensive exploration of these decision boundary intersections represents an intriguing future direction, it leads to the Boundary Value Problem that exceeds the scope of the present study. Nevertheless, the identification of these overlapping regions highlights the complexity of the neural network's decision-making process and presents a potential area of vulnerability, particularly in scenarios where precise decision-making is critical.

A way that these boundaries could be uses is after a Net is trained and then checking if there are any regions defined by the NN that are erroneous Examples of this could be when an image of static is defined as a "3" or the changing of a single isolated pixel's value the NN's prediction. This specific idea is explored later in this section and can be viewed in figure 7.

Another intriguing observation was the occurrence of dormant neurons, i.e., neurons that never activate, even within an accurately performing neural network. This phenomenon was especially evident during our human testing, where

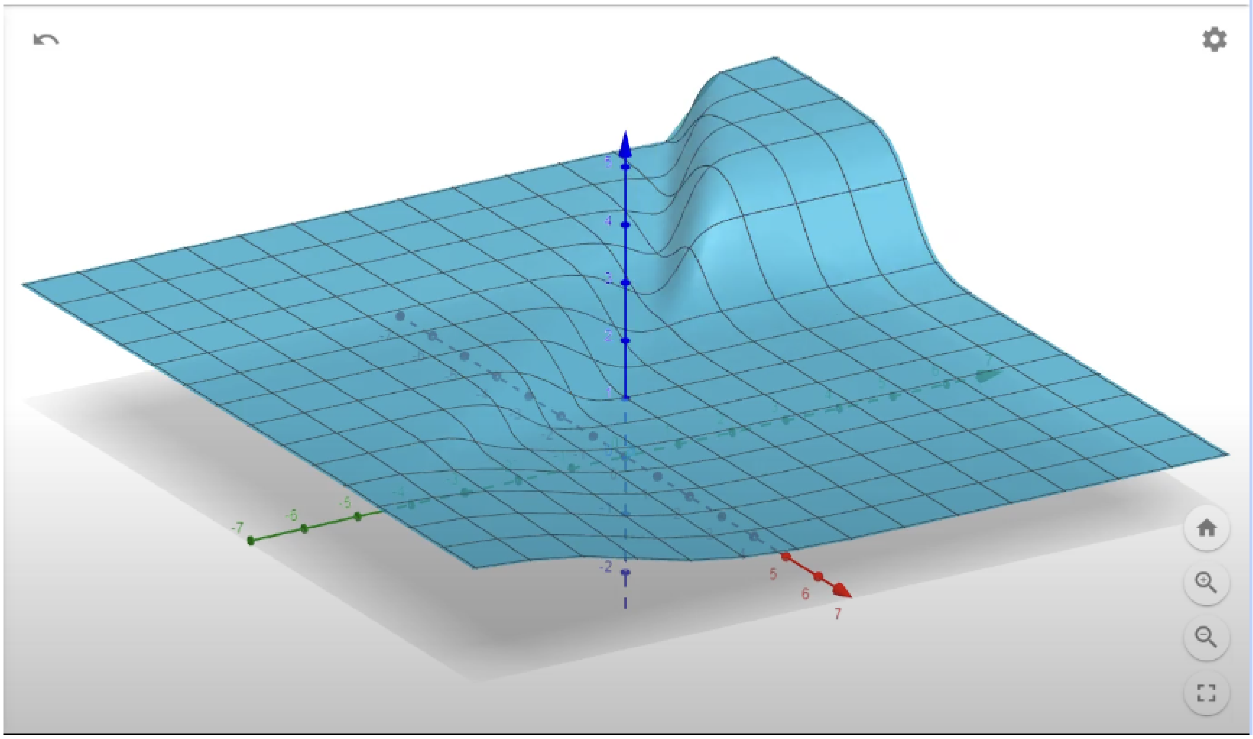

Figure 3: A possible shape of one neuron output given a NN architecture of 2 input neurons, 1 layer hidden layer with 3 output neurons, fully connected and using a Sigmoid activation function.

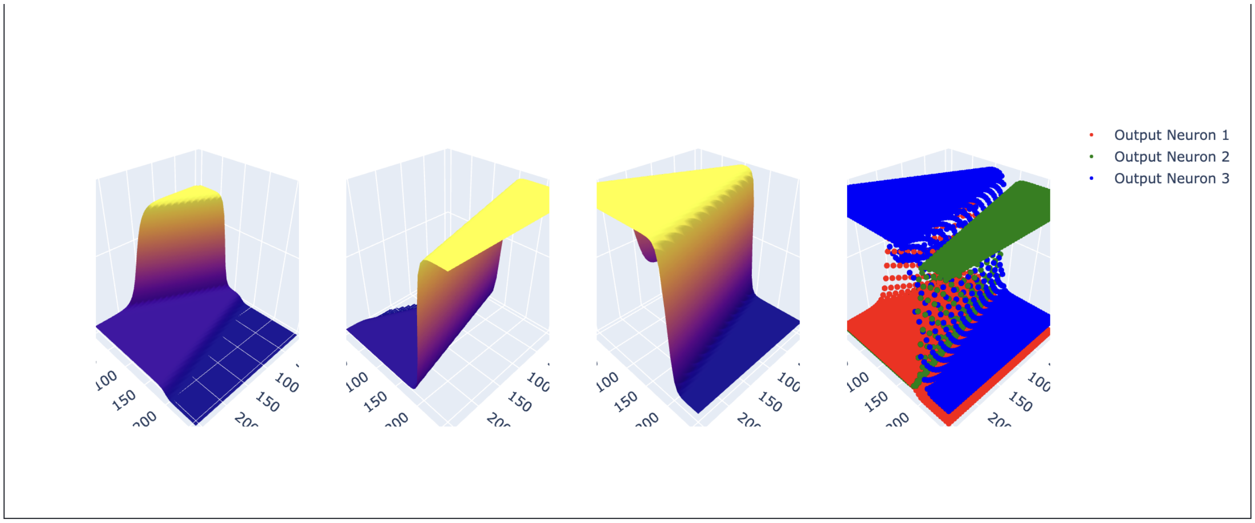

Figure 4: The S1 data set when mapping the two input values to their output the following geometries are generated

it became apparent that a human would not categorize an image as "White" unless the image was entirely white. Interestingly, these dormant neurons are almost or completely inactive. This means that even if the necessary conditions are met the NN will still not predict these outputs.

An additional experiment involved training a Gaussian Naive Bayes model on the MNIST data set. This model achieved a 52 percent accuracy. By traversing every training image and changing a pixel value by ± 1 we were able to pinpoint an erroneous boundary condition. By varying this pixel's value by 1 the model would alter it's class prediction. What was intriguing was that these minor alterations were indistinguishable to the human eye. This highlights the potential

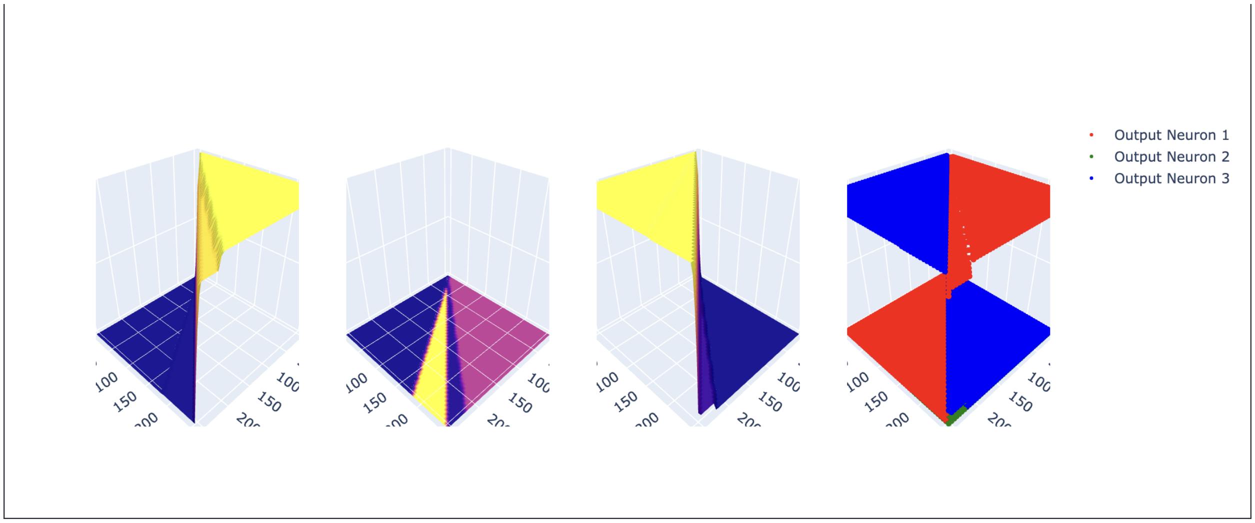

Figure 5: The S2 data set when mapping the two input values to their output the following geometries are generated.

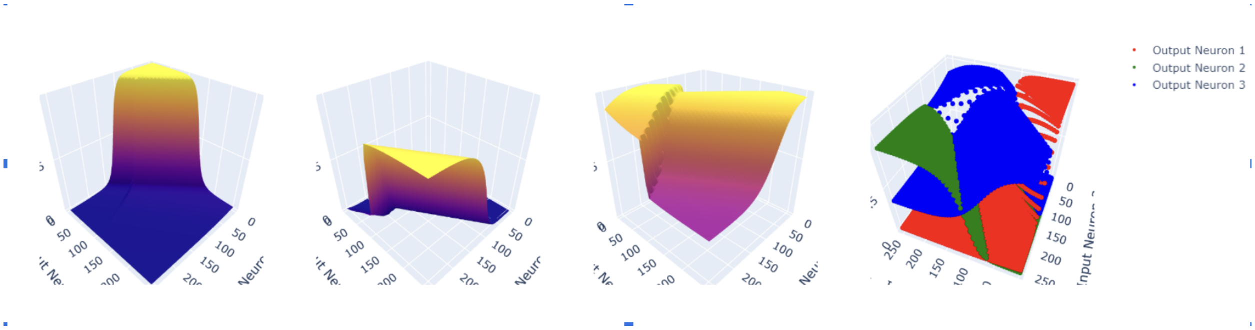

Figure 6: The S3 data set when mapping the two input values to their output the following geometries are generated.

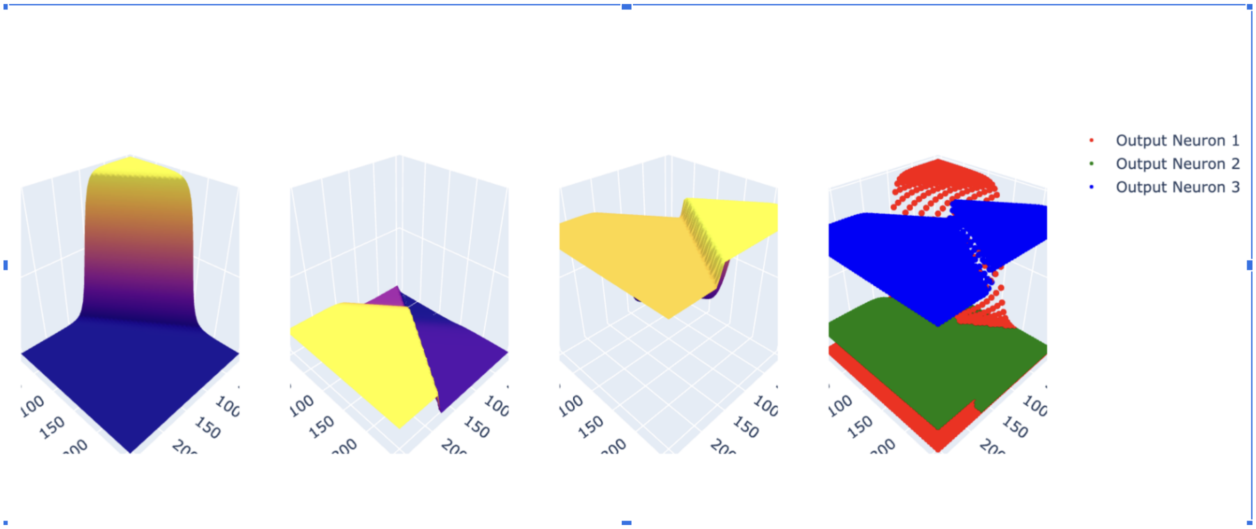

Figure 7: The S4 data set when mapping the two input values to their output the following geometries are generated

for neural networks to perceive subtle differences that humans might miss. This also highlights how NN's can be susceptible to noise and that boundary conditions exist even in NN's that are preforming more complicated tasks.

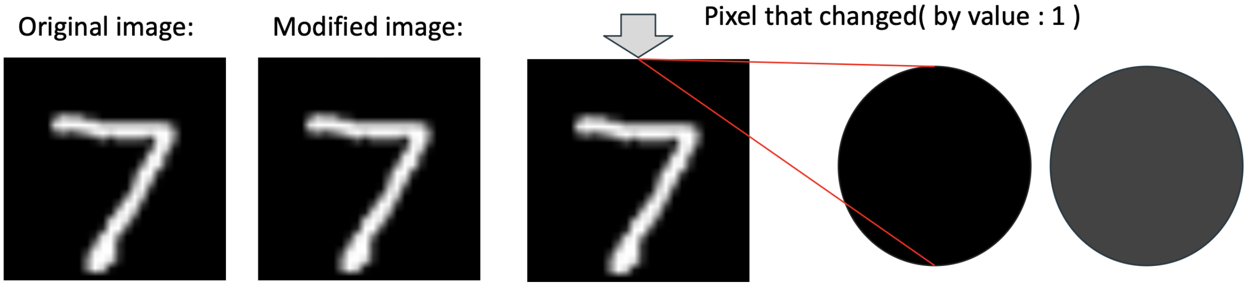

Figure 8: This figure demonstrates that a boundary condition was crossed by changing a single pixel by plus or minus 1

8 Conclusion

This study has undertaken the challenging task of visualizing the minima reached by a neural network during its training process. Our experimental approach, which involved engaging the network with a variety of mathematically defined tasks, allowed us to gain a more intimate understanding of the network's learning and decision-making processes.

Through our experiments, we observed interesting phenomena, such as the overlapping decision planes of individual neurons and the occurrence of dormant neurons in an otherwise accurately performing network. These observations not only highlighted the complexity of neural networks but also pointed to potential limitations and vulnerabilities.

However, our study is only the first step in a broader endeavor. The complexities and high dimensionality of neural networks mean that there are numerous aspects yet to be explored. In particular, the intersection points of decision boundaries, which represent potential regions of ambiguity in the network's decision-making process, are a compelling area for future investigation.

One of the major directions for future work would be to develop algorithms capable of efficiently solving the boundary value problems associated with these high-dimensional decision boundaries. Such an endeavor could lead to improved insights into the functioning and limitations of neural networks, and potentially contribute to the development of more robust and reliable AI systems.

Furthermore, investigating the role of dormant neurons could reveal new information about the learning process of neural networks. Understanding why these neurons remain inactive and what implications this has for the performance of the network can open new avenues for enhancing the efficiency and effectiveness of neural networks.

In conclusion, while this study has made important strides in visualizing and understanding the minima of neural networks, there is much terrain yet to be covered. We believe that the insights gained from this study will serve as a valuable foundation for future research in this domain.

References

[1] W. J. Scheirer, A. Rocha, A. Sapkota, and T. E. Boult, "Toward Open Set Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 7, pp. 1757-1772, July 2013.

[2] D. H. Wolpert, "The Lack of A Priori Distinctions Between Learning Algorithms," Neural Computation, vol. 8, no. 7, pp. 1341–1390, Oct. 1996.

[3] M. Unser, "Spline: A Perfect Fit for Signal and Image Processing," IEEE Signal Processing Magazine, vol. 16, no. 6, pp. 22-38, Nov. 1999.

[4] D. Balduzzi, M. Frean, L. Leary, J.P. Lewis, K.W. Ma, and B. McWilliams, "The Shattered Gradients Problem: If resnets are the answer, then what is the question?" in Proc. 34th International Conference on Machine Learning, Sydney

[5] Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In Computer Vision – ECCV 2014 (pp. 818-833). Springer International Publishing.

[6] Tzeng, F.-Y. and Ma, K.L. (2005). Opening the Black Box - Data Driven Visualization of Neural Networks. In Proceedings of the IEEE Visualization Conference (pp. 383-390).

[7] Balestriero, R., Pesenti, J., & LeCun, Y. (2021). Learning in High Dimension Always Amounts to Extrapolation.

[8] Rahul Parhi and Robert D. Nowak. "What Kinds of Functions Do Deep Neural Networks Learn? Insights from Variational Spline Theory." SIAM Journal on Mathematics of Data Science 4.2 (2022): 464-489.

[9] Randall Balestriero and Richard Baraniuk. "A Spline Theory of Deep Learning." Proceedings of the 35th International Conference on Machine Learning, vol. 80, (2018): 374-383.

[10] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).