EXPLAINABLE ARTIFICIAL INTELLIGENCE: REAPING THE FRUITS OF DECISION TREES

Anonymous authors
Paper under double-blind review

ABSTRACT

The recent push for explainable artificial intelligence (XAI) has given rise to extensive work toward understanding the inner workings of neural networks. Much of that work, however, has focused on manipulating input data feeding the network to assess their effect on network output. It is shown in this study that XAI can benefit from investigating the network node, the most fundamental unit of neural networks. Whereas studies on XAI have mostly benefited from a focus on manipulating input data, assessing patterns in node weights may prove equally beneficial, if not more significant, especially when realizing that weight values may not be as random as previously thought. A manipulated, a contrived, and a real dataset were used in this study. Datasets were run on convolutional and deep neural network models. Node rank stability was the central construct to investigate neuronal patterns in this study. Rank stability was defined as the number of epochs wherein nodes held their rank in terms of weight value compared to their rank at the last epoch, when the model reached convergence, or stability (defined in this study as accuracy ≥ 0.90). Findings indicated that neural networks behaved like a decision tree, in that rank stability increased as weight absolute values increased. Decision tree behavior may assist in more efficient pruning algorithms, which may produce distilled models simpler to explain to technical and non-technical audiences.

1 INTRODUCTION

With the ubiquitous and rapid emergence of machine learning applications in our society, the demands for explainable artificial intelligence (XAI) have been growing equally fast, if not faster (Goodman & Flaxman, 2017; Hoofnagle et al., 2019). Methods employing AI are becoming more accurate in informing decision-makers in government, science, and industry (Ahmed et al., 2022), calling for similar accuracy in explaining outcomes from such methods. Explainability, however, is far from a precise term, requiring context from which to derive its meaning (Miller, 2019), which in turn is contingent on target audience (Miller et al., 2017; Barredo Arrieta et al., 2020; Dazeley et al., 2021). Regardless of audience, understanding the inner workings of neural networks may be advanced if a bottom-up, neuron-focused approach is undertaken, as the ultimate construct determining network outcomes is the neuron.

Understanding the neuronal patterns as a function of input may better assist in defining what neural networks learn and how. Knowledge of the relational link between model input and output may help in the design of better architectures, assist in fine tuning hyperparameters, or create more efficient and safer algorithms (Bhatt et al., 2020; Ivanovs et al., 2021; Quinn et al., 2022), all in the interest of the continued improvements in efficiency of AI systems. In this study, an attempt is made to better understand neural networks by examining neurons. By investigating nodes during network training, patterns may emerge to assist developers, users, and decision makers to better understand model outputs specific to their field.

To achieve the above goal, the remainder of this paper is organized as follows. The next section will cover studies attempting to shed light on black-box models through assessment of neurons. The third section will cover the methodology used in this study toward attempting a contribution to the field of XAI by investigating neurons. The fourth section will cover this study's findings. The last

section will bring together the study by discussing the relevance of the findings herein and proposing lines of future research.

2 Related Work

Most of the effort in better understanding neural network models has been focused on manipulating input data (Bach et al., 2015; Ribeiro et al., 2016a;b; Lundberg & Lee, 2017; Tan et al., 2018; Främling et al., 2021; Ivanovs et al., 2021), with less efforts devoted to a focus on the lower levels, the neurons, of such models. Studies toward more closely assessing the effects of neurons may be approached theoretically or empirically. From the theoretical standpoint, an early study by Goodfellow et al. (2009) assessed network input data representation evolution as such data propagated in the network, showing that deep neural networks progressively became invariant to input transformations while data moved through layers. In a following related study, Montavon et al. (2011) showed that representations of input data became increasingly robust and simpler as data propagated through network layers. Increasing the number of layers yielded more accurate representations at layers away from the input layers. Group theory has also been used in assisting understanding of high-order representations in neural networks (Paul & Venkatasubramanian, 2014). In this latter study, they show that groups of neurons encode features in the input data, and that group size is proportional to feature complexity. Using decision trees has been used to explain how neural networks make decisions when classifying images by selecting a probability distribution over all input classes (Frosst & Hinton, 2017).

Empirical approaches, whereby assessing how perturbations in the input data are reflected in neuron activation patterns, have been a more common choice in XAI research (Rafegas et al., 2020; Ivanovs et al., 2021). One of the earliest approaches for understanding intermediate levels in a neural network was provided by Zeiler & Fergus (2014), who used ablation, an image perturbation method, to assess performance contribution of network layers. A following study applied a generative deconvolution network using neuron activation as the internal network functioning to show realistic object representations, beyond image classification, allowing generalization by the network to unseen data (Dosovitskiy et al., 2015). A similar study by Aubry & Russell (2015) indicated neuron activation changes when images with different scene factors were presented to a CNN. More recently, Fong & Vedaldi (2017) used interpretable perturbations, grounded in theory, providing a framework to identify the relationship between image components and classification outcome. Yang et al. (2021), developed a method to link image input area to classification output without needing to understand any model architectural details. One may conclude that image perturbation studies suggest specific neurons are being responsible for different attributes of the input image.

Still on empirical XAI approaches, convergent learning, whereby separately trained networks converge toward similar spaces, has been shown to occur with neuron one-to-one and many-to-many matching approaches between networks Li et al. (2015). Spatio-chromatic representations of image color have also been shown to be reflected in neuron activation patterns (Rafegas et al., 2020). Using semantic dictionary mapping between neuron activation patterns and input images, Olah et al. (2017; 2018) provide an approach to understanding which image regions are responsible for determining neuron activation. Network dissection has also been successfully used in XAI using neuron activation. Using such method, Bau et al. (2018) and Zhou et al. (2019) used activation on multilabel images with distinct colors, materials, and shapes to link image characteristics and neuron activation. More recently, using a combination of network dissection and expert annotation in mammalogy, Wu et al. (2021) was able to better explain model predictions.

The studies above show a significant potential in focusing on the lowest level of network models to better understand how inputs generate outputs in neural networks, providing ample justification for focusing research efforts on nodes toward the advancement of XAI. Those studies, however, did not consider in great detail the neurons, the most fundamental level of neural networks. By investigating node weight patterns according to the methodology below, this work aims to close that gap, furthering our understanding of the relationship between input data and the outcomes of machine learning applications currently so much prized in the decision-making process of most institutions and organizations in our society.

3 METHODS

3.1 DATASETS

A processed, contrived, and real dataset were used in this study. The processed dataset (PR) was a reduced version of the MNIST handwritten dataset. The MNIST dataset was reduced to 100 images per category to keep accuracies from reaching high values early during training. Avoidance of high early accuracy was to enable investigation of patterns in weight value progression during training, as per experiments below. Reduction of the MNIST dataset was achieved by randomly selecting the above number of images per category.

The real dataset (SC, for scales) was used to provide an example of how XAI, especially with the approach used in this study, might be applied in practice. The real dataset was for grayscale images of Gulf menhaden (*Brevoortia patronus*) scales read independently by three biologists to infer fish age, a key determinant for sustainably managing fisheries resources (Schueller et al., 2021). Classes for this dataset were 0-4, the possible observed ages for the target fish. The number of images within each class was 530, of 100 by 100 pixels each.

The contrived dataset (CT) was generated to have better control and understanding of model output, especially the number of epochs before stabilization, as explained below. Having better control of model output might enable finer and more accurate study findings and interpretation. The contrived dataset comprised of grayscale images with 10 shades of gray, producing 10 classes, which formed the inputs to neural network models. The classes for this dataset were generated by sampling from a Gaussian distribution with means of 20, 50, 75, 100, 125, 150, 175, 200, 225, and 240 to mimic pixel depth for grayscale images. The means selected for classes were chosen to ensure an even spread of shades of gray among classes. The standard deviation was 1,000 across all classes, which allowed model convergence within an adequate number of epochs to conduct study experiments. Data points that were below zero or above 255, the range for grayscale image shades of gray, were forced to take the minimum or maximum value, respectively, of that range. The number of generated images within each class was 50, each of size 100 by 100 pixels.

3.2 EXPERIMENTS

Experiments were done to investigate patterns in the progression during training of node weights according to AI models, datasets, and epochs. Epochs were used as the time surrogate in assessing training weight progression. The models tested were deep neural networks (DNN) and convolutional neural networks (CNN). The DNN model architecture consisted of a flattened input layer, six fully connected layers of 30 x 30 nodes, and the output softmax layer. The CNN model comprised of an input, a fully connected, and an output set of layers. The fully connected and output layers were as in the DNN model. The CNN input layer consisted of a sequence of three convolution layer (3 x 3 kernel) and max pooling layers, followed by another convolution layer, and ending with a flattened layer. The flattened CNN layer served as the input to the fully connected layer. To keep comparisons consistent, a ReLU activation function, Adam optimizer, He-Normal initializers, batch size of 32, and the learning rate of 0.001 were used throughout.

Images in the datasets above were fed one at a time into the input layer of the above models. Images were flattened and padded to keep size consistence prior to model input. Padding was done by addition of extra pixels of value 0, mimicking an empty, black background. Each image pixel corresponded to a node in the input layer to the model tested.

Both DNN and CNN models were run toward stabilization on the three datasets above in replicates of 50. Each replicate was run for 150 epochs. Model runs that did not converge to at least 0.90 percent accuracy were re-run for inclusion in analysis. After each epoch, the current model accuracy and the node weights from the layer before that for the softmax activation were extracted for node pattern analysis (focus layer henceforth). Node analysis consisted of an assessment of whether the progression of nodes from the initial run to stabilization could be modeled by a decision tree, potentially identifying the most significant nodes in determining model accuracy early in the training process.

3.3 DATA ANALYSIS

Analyses in this work were based on the construct of node rank stability, herein defined as the number of epochs where individual nodes of the focus layer retained their weight rank from that of the stable model. To determine rank stability, all nodes from the focus layer of the last epoch (stable model) were ranked according to weight magnitude and followed backward to the first epoch. The number of epochs where individual nodes retained their rank was defined as the rank stability for that node. Rank stability, therefore, ranged between one and 150, the maximum number of epochs.

To test for neural network decision tree behavior, the statistical significance of rank stability was assessed. Decision tree behavior was assessed with a generalized linear model (GLM) using rank stability as the response variable with a quasi-Poisson link function. The predictors for the GLM were the model type, of levels DNN or CNN, the datasets, of levels CT, PR, and SC, and the group to where node weight magnitude was assigned. Node weight group assignment was done because model stabilization is more likely due to multiple, rather than individual nodes (Li et al., 2015). In light of minimizing subjectivity, node groups were formed using Jenks Natural Breaks method (Jenks & Coulson, 1963), which minimizes within-class variation and maximizes among-class variation. The number of breaks for determining node group assignment was chosen as four, namely highest positive, lowest positive, lowest negative, highest negative values. The Tukey's Honest Significance post-hoc GLM was used to infer within-mean differences for statistically significant factors.

An additional test was performed assessing the relationship between the number of stable nodes and training time. The relationship was tested in two steps. The first step consisted of using a linear regression to estimate the slope of the relationship between epochs, the independent variable surrogate for time, and the number of stable nodes in the focus layer for each replicate as above. Stable nodes were defined as nodes that were rank-stable at a given epoch, i.e., nodes that retained their row and column position from that of the last epoch. For each epoch, stable nodes were counted and used as the response variable in the regression analysis. Because the count of stable nodes with epoch tended to follow an exponential curve, the response variable was logarithm-transformed prior to estimating regression parameters. The second step in analyzing node progression was to examine the slopes of the regression analysis obtained in the first step. An analysis of variance (AOV) using as factors model and dataset, and the regression slopes as the response variable was conducted.

3.4 DATA VISUALIZATION

To further assess decision tree behavior, two visualization approaches with reduced number of nodes were developed for the focus layer according to the above experiment. Because of the large number of nodes to visualize (900/epoch), a node grouping approach was used to display rank stability patterns. Node grouping was based on their weight value. Node group membership was determined using the Jenks method above, but with varying number of breaks. Breaks numbering 50, 200, 400, 600, and 800 for groupings were used to provide a wide contrast for displaying rank stability as a function of Jenks breaks.

For the first visualization approach, only three groups were retained for display after groups were generated based on the breaks above. The first group was for the nodes belonging to the group of lowest weights, i. e., highest negative. The second group was of the nodes with intermediary weights, i. e., around zero. Finally, the third group consisted of the nodes with the highest weight values, i. e., highest positive. The three groups for display were generated for each combination of Jenks number of breaks, models, and datasets (30 displays). Within each such groups, the mean rank stability was graphed. The mean and standard deviation of the weight values, as well as the number of nodes, within each group were also displayed.

The second approach displayed the within-group mean rank stability for nodes for all groups, but only for breaks numbering 50, 400, and 800. This latter approach allowed for discovery of finer patterns in rank stability as a function of Jenks number of breaks, models, and datasets. A second-order polynomial was fit to the rank stability data as an aid for visualizing the shape of the relationship between rank stability and number of Jenks breaks for each combination of the models and datasets discussed above.

4 RESULTS

All experimental runs converged after 150 epochs, many of which achieved accuracies well above 0.90. Average rank stability by model and dataset ranged between 3 and 95, with associated standard deviations of 3.66 and 37.57 (Table 1), indicating a high consistency in node rank stability. For runs with high rank stability, the assertion of model training following a decision tree behavior is made stronger. A high rank stability, such as for all CNN and most DNN runs (Table 1), is suggestive of the presence of nodes, root nodes as an analogy, determining model outcomes early during training.

Providing further details on decision tree behavior trends, node rank stability averages within model and dataset were always highest for groups 1 and 4, and lowest for groups 2 and 3, indicating that stability indeed is a function of weight absolute values and, thus, node significance in determining model performance. Stability rank overall, however, showed some variation. For CNN models run on the PR and SC datasets, nodes in groups 1 and 4 kept their rank for longer. For the CNN model runs on the CT dataset, groups 1 and 4 also showed highest rank stability, although those values were lower than CNN runs on the other datasets. For the DNN model runs across datasets, the pattern of higher rank stability for groups 1 and 4 generally held true (Table 1).

Table 1: Summary statistics of node rank stability for experiments using CNN and DNN models on a contrived (CT), processed (PR), and fish scale (SC) grouped dataset; groups were according to 4 Jenks Natural Breaks classes; MRS: mean rank stability, SSD: rank stability standard deviation.

MODEL	DATASET	GROUP	MRS	SSD
CNN	CT	G1	68	38.7
CNN	CT	G2	58	39.0
CNN	CT	G3	58	39.8
CNN	CT	G4	69	39.5
CNN	PR	G1	90	44.8
CNN	PR	G2	72	43.5
CNN	PR	G3	70	43.5
CNN	PR	G4	87	44.8
CNN	SC	G1	95	37.6
CNN	SC	G2	79	41.1
CNN	SC	G3	79	40.8
CNN	SC	G4	92	38.4
DNN	CT	G1	25	30.3
DNN	CT	G2	19	24.1
DNN	CT	G3	17	22.5
DNN	CT	G4	23	28.2
DNN	PR	G1	33	33.1
DNN	PR	G2	20	19.0
DNN	PR	G3	18	18.3
DNN	PR	G4	34	32.3
DNN	SC	G1	5	9.8
DNN	SC	G2	3	3.6
DNN	SC	G3	3	3.7
DNN	SC	G4	5	10.5

The GLM results showed model, datasets, and groups to be highly significant predictors (Table 2). The dataset factor had the highest influence on the GLM results (as per F-values), followed by the model, and grouping factors, indicating that rank stability is highly dependent on the variations inherent in model inputs. Post-hoc analysis indicated that the mean rank stability difference between groups 1 and 4, and groups 2 and 3 were low, whereas the difference between groups 1 and 2, and 3 and 4 were high (Table 3), suggesting similarity in terms of rank stability between groups of low differences.

The number of stable nodes at the focus layer as a function of epoch followed a linear relationship for CNN and an exponential for DNN models. For the DNN model run on the SC dataset, the slope of

Table 2: Results for the generalized linear model testing node rank stability for experiments using CNN and DNN models on a contrived (CT), processed (PR), and fish scale (SC) grouped dataset; DF: degrees of freedom, SSQ: sums-of-squares; groups were according to 4 Jenks Natural Breaks classes.

FACTORS	DF	SSQ	F-VALUE	p-VALUE
Group	1	1,073	38	<0.01
Input Data	1	49,102	1,719	<0.01
Model	1	5,527,431	193,454	<0.01
Residuals	269,996	7,714,400		

Table 3: Tukey's Honest Significance Difference post-hoc results of node rank stability for experiments using CNN and DNN models on a contrived (CT), processed (PR), and fish scale (SC) grouped dataset; groups were according to 4 Jenks Natural Breaks classes.

	DIFFERENCE	LOWER	UPPER	p-ADJUSTED
Groups				
G2-G1	-11.32	-11.81	-10.84	<0.01
G3-G1	-12.02	-12.51	-11.54	<0.01
G4-G1	-1.24	-1.78	-0.69	<0.01
G3-G2	-0.7	-1.13	-0.28	<0.01
G4-G2	10.09	9.6	10.58	<0.01
G4-G3	10.79	10.3	11.28	<0.01
Datasets				
CT-PR	-10.09	-10.46	-9.72	<0.01
SC-PR	-6.98	-7.36	-6.61	<0.01
SC-CT	3.11	2.74	3.48	<0.01
Models				
CNN-DNN	58.52	58.78	58.27	<0.01

the exponential relationship grew rapidly late during training, indicating model stabilization toward the end of training. The run for the CNN model on the SC dataset showed a weak indication of a plateau, as it grew rapidly early in training and slowed down at about epoch 50 (Figure 1).

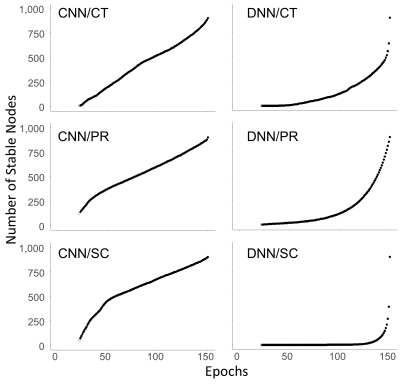

Figure 1: Relationship between the mean number of stable nodes in the focus layer and epoch; CT: contrived, PR: processed, SC: scale datasets.

The results following AOV analysis on the slope of the regression of epochs on the logarithm of stable node counts showed to be highly significant for the factors model and dataset. As per the F-value, the factor dataset was more influential in determining the slope than was the factor model

(Table 4), a parallel result from the GLM analysis. Post-hoc AOV results showed that all datasets were significantly different from each other, which was also the case for the factor model (Table 4).

Table 4: Results for the AOV model testing node progression toward stabilization for experiments using CNN and DNN models run on a contrived (CT), processed (PR), and fish scale (SC) dataset; DF: degrees of freedom, SSQ: sums-of-squares.

FACTORS	DF	SSQ	F-VALUE	p-VALUE
Model	1	0.00081	5.87	<0.01
Dataset	2	0.01047	37.73	<0.01
Residuals	296	0.04107		

Displays for node stability using a range of Jenks breaks confirmed the AOV findings above (Figures 2 and 3). For the first visualization approach, rank stability was highest for CNN models run on the SC dataset, including nodes with lesser significance (close to zero value). The nodes of low significance, however, were always of lower rank stability. Stability rank for the most significant nodes (highest absolute values) within the groups of negative and positive weights was generally similar. The rank stability tended to be similar across the range of number of Jenks breaks (Figure 2).

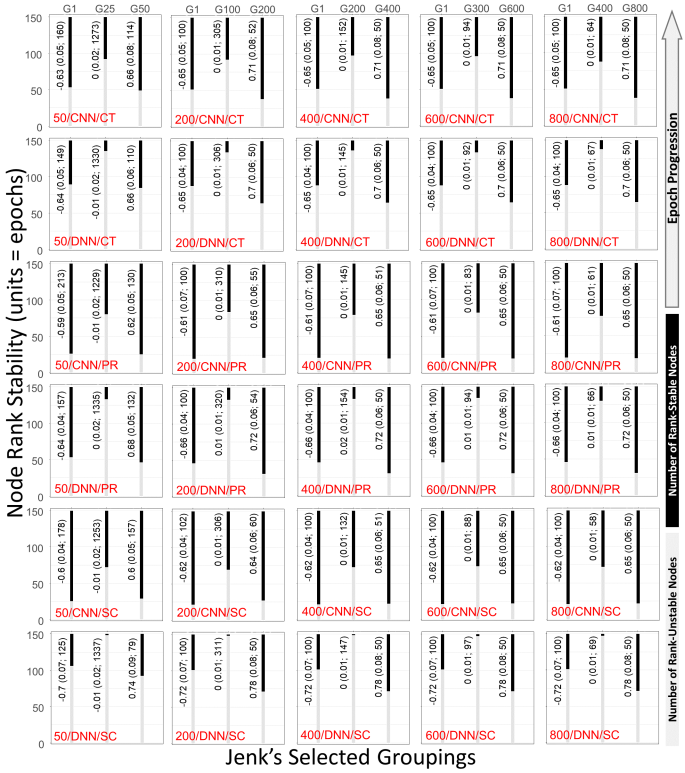

Figure 2: Mean node rank stability as a function of Jenks groupings, model, and dataset; red labels indicate number of Jenks breaks for group creation/model/dataset; CT: contrived dataset, PR: processed dataset, SC: scale dataset; numbers in vertical orientation are mean±95 percent confidence intervals of node weights and number of nodes within Jenks groups; only first, middle, and last group from Jenks grouping shown.

For the second visualization approach, rank stability was low for most of the intermediate nodes after groups constructed using Jenks breaks. In general, CNN models produced higher rank stabilities at the extreme groups, that is for the highest negative and highest positive weights, than did DNN

models. The CNN models also displayed higher overall rank stabilities. The DNN model run on the SC dataset produced more varying mean rank stability across groups, regardless of the number of Jenks breaks used (Figure 3).

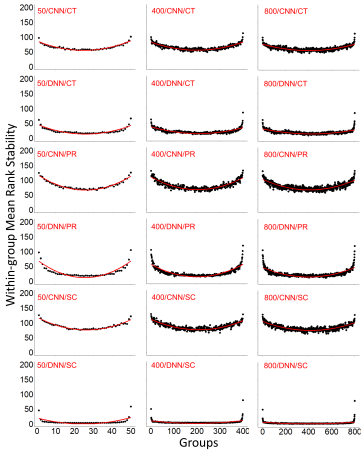

Figure 3: Mean node rank stability as a function of Jenks groupings, model, and dataset; red labels indicate number of Jenks breaks for group creation/model/dataset; CT: contrived dataset, PR: processed dataset, SC: scale dataset; all groups formed by Jenks breaks shown.

5 Discussion

The evidence in this study using the concept of rank stability supports the contention that progression in model training operates as a decision tree, as opposed to following an unpredictable, random pattern. As indicated by node weight value groupings, groups of nodes with high mean absolute values tended to retain their rank for longer, implying that their significance in determining model output was also retained for longer. Such nodes are analogous to the concept of decision tree roots. The timing when those weights and their ranks were established might have varied among individual runs but did show a clear pattern according to groups. This lends support to the existence of a critical period when training performance is determined (Golatkar et al., 2019). Establishing that critical period may be useful in determining network performance and pruning. The critical period might be best established by visualizing the network as a decision tree.

The strength of the behavior as a decision tree using rank stability was not observed to be consistent throughout models and datasets. For CNN models, rank stability tree behavior was more pronounced. Moreover, tree behavior was dependent on the dataset used. Even though rank stability was always higher for the groups 1 and 4, groups with the higher absolute weight values, and lower for groups 2 and 3, models and input data were critical in determining stability. The GLM results showed that the input data not only is important but also had the highest influence in determining rank stability. This may be due to the higher variation inherent in datasets, as compared to the intrinsic randomness during model training. Models, in turn, showed an intermediate influence, followed by the grouping factor.

The subsequent analyses using different groupings and slopes of regression of number of rank-stable nodes on training epochs confirmed the relative importance among model, datasets, and Jenks groupings (AOV only). Invariably, rank stability was highest for nodes with either high or low weight values and lowest for intermediate values. Moreover, model stabilization was due to a few, highly significant nodes, rather than a more evenly distributed node weight in determining model stabilization. The grouping factor for the GLM and AOV analyses above, however, is the indicator of decision tree behavior, because it is the factor that reflects node weights.

Even though the grouping factor showed the weakest effect on indicating that model training follows a decision tree behavior, it is the only one that cannot be controlled by the user. Models can be chosen and their configuration adjusted. Datasets can be manipulated and cleaned off of suspect data points.

Once model choices and data processing are completed, however, the user is left with the vagaries of randomness in the algorithm responsible for determining which nodes or groups thereof influence model stabilization. Using rank stability to prune a model cannot, therefore, be conducted a priori, because the weights early during training are as of yet not stable. Periodic checking of rank and rate of change of weight values may be necessary during training.

The number of stable nodes by epoch tended to be linear for CNN models and follow an exponential for DNN models, indicating that the addition of influential nodes grows more evenly during training for the former model. For DNN models, however, this indicates that influential nodes are added late. It is also worth noticing that fewer influential nodes were observed for DNN models than for CNN models. This finding may be critical in understanding model output. If fewer nodes are significant, linkages between model weight patterns and patterns of model input data can be simplified and possibly made more apparent.

5.1 Conclusions and Future Work

In general, progression from model initial conditions to stabilization followed a decision tree behavior, with the most influential nodes, those with either highly positive or highly negative weights, being set first. Such nodes became significant in determining model stabilization early during training, whereas the least significant ones tended to fix their rank at later epochs. This work also showed that the progression toward model stabilization using rank stability as a metric was not always linear. The shape of the relationship between the nodes with fixed rank and epoch was found to be highly exponential in some cases, suggesting that model stability occurred later during training. For runs showing exponential trends, using the findings of this study to lower training time through tasks such as pruning may not be as optimal as one would expect. Regardless of findings, however, this study points to the possibility of making training more efficient and models more transparent. Expanding on this study's findings, therefore, may only further add to the potential benefits of this work.

The success of neural networks is undisputed. With that success, however, comes computational power and model complexity. Neural networks are increasingly becoming overparameterized (Denil et al., 2013; Sze et al., 2020; Yeom et al., 2021), which may hinder their explainability. Retaining the most significant neural network units, therefore, becomes necessary in light of network transparency. Filtering out nodes early may also make training more efficient and simplify networks to the extent that relationships between inputs and outputs are better understood, making networks more transparent. An additional benefit from this study may be in reducing computational costs of running neural networks. Computational costs may be reduced through hardware choices and algorithm adjustments (Sze et al., 2020). Findings from this study may assist in development of algorithms based on early stopping during training, lowering the time for learning completion without compromising output accuracy. One other example where the findings of this study may benefit is for network specialization (Balemans et al., 2022), whereby networks are compressed for subtask specialization. Early pruning may aid in compression efficiency, while retaining network classification accuracy. Another area of interest where this work may assist is in network information retention (Ede et al., 2022), which aims at minimizing loss of previous learning when new data is added during training. By retaining and possibly fixing weights with high rank stability, one might be able to minimize learning losses.

The studies above may benefit from the findings herein, making extending the results of this study a potential additional benefit. As such, one may consider additional model layers for experimentation. The results of this study apply only to the last layer before the softmax of CNN and DNN models. Assessing other layers might provide a better picture of how a network learns. Adding additional layers from the fully connected layers of CNN and DNN models might offer important insights toward transparency that supplements this work's contribution, as well as that from previous work focused on model input layers. An additional important research area may be investigation on how to implement early pruning based on this study's findings. Detecting significant nodes might have to involve monitoring of each node rank and rate of change in weight values. Such monitoring may assist in detecting nodes with high rank stability, which can be fixed while pruning those with low rank stability. Future work extending the results of this study may not only offer better, more effective models, but also, possibly more importantly, aid in closing the gap between high-level and low-level model explanations, hopefully bringing black-box models closer to their white-box counterparts.

REFERENCES

Imran Ahmed, Gwanggil Jeon, and Francesco Piccialli. From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where. *IEEE Transactions on Industrial Informatics*, 18(8):5031–5042, August 2022. ISSN 1941-0050. doi: 10.1109/TII.2022.3146552. Conference Name: IEEE Transactions on Industrial Informatics.

Mathieu Aubry and Bryan Russell. Understanding deep features with computer-generated imagery, June 2015. URL http://arxiv.org/abs/1506.01151. arXiv:1506.01151 [cs].

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, July 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130140. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140. Publisher: Public Library of Science.

Dieter Balemans, Philippe Reiter, Jan Steckel, and Peter Hellinckx. Resource efficient AI: Exploring neural network pruning for task specialization. *Internet of Things*, 20:100599, November 2022. ISSN 2542-6605. doi: 10.1016/j.iot.2022.100599. URL https://www.sciencedirect.com/science/article/pii/S2542660522000841.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.12.012. URL https://www.sciencedirect.com/science/article/pii/S1566253519308103.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks, December 2018. URL http://arxiv.org/abs/1811.10597. arXiv:1811.10597 [cs].

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 648–657, New York, NY, USA, January 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3375624. URL http://doi.org/10.1145/3351095.3375624.

Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299:103525, October 2021. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103525. URL https://www.sciencedirect.com/science/article/pii/S000437022100076X.

Misha Denil, Babak Shakibi, Laurent Dinh, Marc' Aurelio Ranzato, and Nando de Freitas. Predicting Parameters in Deep Learning. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/hash/7fec306d1e665bc9c748b5d2b99a6e97-Abstract.html.

Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1538–1546, June 2015. doi: 10.1109/CVPR.2015.7298761. ISSN: 1063-6919.

Sami Ede, Serop Baghdadlian, Leander Weber, An Nguyen, Dario Zanca, Wojciech Samek, and Sebastian Lapuschkin. Explain to Not Forget: Defending Against Catastrophic Forgetting with XAI. In Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, and Edgar Weippl (eds.), *Machine Learning and Knowledge Extraction*, Lecture Notes in Computer Science, pp. 1–18, Cham, 2022. Springer International Publishing. ISBN 978-3-031-14463-9. doi: 10.1007/978-3-031-14463-9_1.

Ruth C. Fong and Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, Venice, October 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.371. URL http://ieeexplore.ieee.org/document/8237633/.

Nicholas Frosst and Geoffrey Hinton. Distilling a Neural Network Into a Soft Decision Tree, November 2017. URL http://arxiv.org/abs/1711.09784. arXiv:1711.09784 [cs, stat].

Kary Främling, Marcus Westberg, Martin Jullum, Manik Madhikermi, and Avleen Malhi. Comparison of Contextual Importance and Utility with LIME and Shapley Values. In Davide Calvaresi, Amro Najjar, Michael Winikoff, and Kary Främling (eds.), *Explainable and Transparent AI and Multi-Agent Systems*, Lecture Notes in Computer Science, pp. 39–54, Cham, 2021. Springer International Publishing. ISBN 978-3-030-82017-6. doi: 10.1007/978-3-030-82017-6_3.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Time Matters in Regularizing Deep Networks: Weight Decay and Data Augmentation Affect Early Learning Dynamics, Matter Little Near Convergence, May 2019. URL http://arxiv.org/abs/1905.13277. arXiv:1905.13277 [cs, stat].

Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring Invariances in Deep Networks. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/hash/428fca9bc1921c25c5121f9da7815cde-Abstract.html.

Bryce Goodman and Seth Flaxman. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine*, 38(3):50–57, October 2017. ISSN 2371-9621. doi: 10.1609/aimag.v38i3.2741. URL https://ojs.aaai.org/index.php/aimagazine/article/view/2741. Number: 3.

Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius. The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, January 2019. ISSN 1360-0834. doi: 10.1080/13600834.2019.1573501. URL https://doi.org/10.1080/13600834.2019.1573501. Publisher: Routledge _eprint: https://doi.org/10.1080/13600834.2019.1573501.

Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, October 2021. ISSN 0167-8655. doi: 10.1016/j.patrec.2021.06.030. URL https://www.sciencedirect.com/science/article/pii/S0167865521002440.

G. Jenks and M. Coulson. Class intervals for statistical maps. *International Yearbook of Cartography*, 3:119–134, January 1963.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent Learning: Do different neural networks learn the same representations? In *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, pp. 196–212. PMLR, December 2015. URL https://proceedings.mlr.press/v44/li15convergent.html. ISSN: 1938-7228.

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, February 2019. ISSN 0004-3702. doi: 10.1016/j.artint.2018.07.007. URL https://www.sciencedirect.com/science/article/pii/S0004370218305988.

Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences, December 2017. URL http://arxiv.org/abs/1712.00547. arXiv:1712.00547 [cs].

Gregoire Montavon, Mikio L Braun, and Klaus-Robert Mueller. Kernel Analysis of Deep Networks. *Journal of Machine Learning Research*, 12:19, January 2011.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature Visualization. *Distill*, 2(11):e7, November 2017. ISSN 2476-0757. doi: 10.23915/distill.00007. URL https://distill.pub/2017/feature-visualization.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The Building Blocks of Interpretability. *Distill*, 3(3):e10, March 2018. ISSN 2476-0757. doi: 10.23915/distill.00010. URL https://distill.pub/2018/building-blocks.

Arnab Paul and Suresh Venkatasubramanian. Why does Deep Learning work? - A perspective from Group Theory. *http://arxiv.org/abs/1412.6621*, December 2014.

Thomas P. Quinn, Stephan Jacobs, Manisha Senadeera, Vuong Le, and Simon Coghlan. The three ghosts of medical AI: Can the black-box present deliver? *Artificial Intelligence in Medicine*, 124: 102158, February 2022. ISSN 0933-3657. doi: 10.1016/j.artmed.2021.102158. URL https://www.sciencedirect.com/science/article/pii/S0933365721001512.

Ivet Rafegas, M. Vanrell, and Luís A. Alexandre. Understanding trained CNNs by indexing neuron selectivity. *Pattern Recognit. Lett.*, 2020. doi: 10.1016/j.patrec.2019.10.013.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-Agnostic Interpretability of Machine Learning, June 2016a. URL http://arxiv.org/abs/1606.05386. arXiv:1606.05386 [cs, stat].

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, August 2016b. URL http://arxiv.org/abs/1602.04938. arXiv:1602.04938 [cs, stat].

Amy M. Schueller, Amanda Rezek, Raymond M. Mroch III, Eric Fitzpatrick, and Alicia Cheripka. Comparison of ages determined by using an Eberbach projector and a microscope to read scales from Atlantic menhaden (Brevoortia tyrannus) and Gulf menhaden (B. patronus). *Fishery Bulletin*, 119(1):21–32, April 2021. ISSN 00900656. doi: 10.7755/FB.119.1.4.

Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. *Efficient Processing of Deep Neural Networks*. Morgan & Claypool Publishers, San Rafael, June 2020. ISBN 978-1-68173-835-2.

Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 303–310, New York, NY, USA, December 2018. Association for Computing Machinery. ISBN 978-1-4503-6012-8. doi: 10.1145/3278721.3278725. URL http://doi.org/10.1145/3278721.3278725.

Jimmy Wu, Bolei Zhou, Diondra Peck, Scott Hsieh, Vandana Dialani, Lester Mackey, and Genevieve Patterson. DeepMiner: Discovering Interpretable Representations for Mammogram Classification and Explanation. *Harvard Data Science Review*, 3(4), October 2021. doi: 10.1162/99608f92.8b81b005. URL http://arxiv.org/abs/1805.12323. arXiv:1805.12323 [cs].

Qing Yang, Xia Zhu, Jong-Kae Fwu, Yun Ye, Ganmei You, and Yuan Zhu. MFPP: Morphological Fragmental Perturbation Pyramid for Black-Box Model Explanations. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 1376–1383, January 2021. doi: 10.1109/ICPR48806.2021.9413046. ISSN: 1051-4651.

Seul-Ki Yeom, Philipp Seegerer, Sebastian Lapuschkin, Alexander Binder, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Pruning by explaining: A novel criterion for deep neural network pruning. *Pattern Recognition*, 115:107899, July 2021. ISSN 0031-3203. doi: 10.1016/j.patcog.2021.107899. URL https://www.sciencedirect.com/science/article/pii/S0031320321000868.

Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pp. 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1. doi: 10.1007/978-3-319-10590-1_53.

Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting Deep Visual Representations via Network Dissection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2131–2145, September 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2858759. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.