Analysis of Activation Function Order in Multi-Activation Neural Network Architectures

Jaskirat Sohal

jsohal@sutdents.kennesaw.edu

https://github.com/JaskiratSohal/Activation-Function-Analysis-of-Multi-Activation-Neural-Networks

Department of Computer Science, Kennesaw State University

Marietta, GA

Abstract

Neural Networks are commonly used to detect and learn the patterns within a dataset by employing activation functions, creating a uniform function based on specific inputs. The introduction of the Multi-Activation layer Neural Networks allows for the sequential distribution of multiple activation functions throughout the model per layer. This paper explores the effects on the performance of a Sequential Multi-Activation model based on the order of the activation functions per layer. The Sequential and recently proposed Multi-Activation Neural Networks were tested using datasets of varying size and distributions of target values, showcasing that altering the order of the activation functions for a model can lead to increased accuracy.

Introduction

Artificial Intelligence is a method of perceiving, analyzing, and extracting hidden information and patterns from datasets, including but not limited to audio, images, text, and videos. Artificial intelligence can be separated into multiple subsets, such as deep learning, neural networks, language processing, machine learning, and robotics. Machine learning can be split into three subcategories. These categories are labeled unsupervised, supervised, and reinforcement learning [1]. Supervised learning occurs when an algorithm attempts to recognize and learn the pattern within the dataset provided. Most used machine learning methods employ deep learning and Neural Network models.

Neural Networks attempt to replicate the biological neurons present in a human brain. Any input or pass-through of information to the neural network can be seen as the dendrites and synapses in action, while the activation functions that process and update the information in the neural networks can be perceived as a neuron firing off an action potential along its axon towards the next [3]. As our neurons develop with actions performed over a long period and become our habits, neural networks learn and improve over multiple iterations while parsing through the dataset provided.

Every unit in a neural network linearly combines the input features and applies activation to create a non-linear mapping for the selected dataset [4]. The networks

learn the pattern in the dataset through multiple iterations and adjust the weight according to the loss by employing backpropagation and gradient descent based on the activation function. Multiple activation functions allow for different implementations of a function, as various functions have different output values.

A wide range of research explores multiple activation functions and their applications. The function's derivative determines how quickly specific Neural Networks can learn the patterns. Application of a particular activation function throughout the entire neural network's layers is commonly practiced. However, certain studies have explored the execution of Neural Networks with different activation functions in different hidden layers [4], [5].

This study showcases the effects of the ordering of the activation function and the performance of a Multi-activation Neural Network model with different optimization algorithms outperforming commonly used machine learning classifiers such as Logistic Regression, Linear SVM, Fine KNN, Weighted KNN, Sigmoid DNN, ReLU DNN, etc., under specific condition. All the models are tested on three different datasets containing binary and multi-classification.

This paper explores the impact of the activation function order present in a Sequential Multi-Activation Neural Network [5] while also utilizing a recently proposed architecture that achieved results comparable with identical uniform architectures that applied a single activation function throughout the network [2].

Related Work

As previously mentioned, a Multi-Activation Neural Network model, referred to as Paper MANN in the figures and tables, has been tested for breast cancer diagnosis using the Adagrad optimizer. The study proposed a Deep Neural Network consisting of 7 total layers, an input layer, a hidden layer consisting of 30 neurons, a dropout layer with a rate of 0.3, a hidden layer consisting of 15 neurons, another dropout layer with a rate of 0.3, a hidden layer consisting of 30 neurons, and an output layer containing a single neuron. Swish, LeakyReLU, ReLU, and Sigmoid are applied to the hidden and final output layers, respectively. The results indicate improved accuracy, F1-Scores, Sensitivity, and Recall compared to traditionally used classification models [5].

Drawing inspiration from the Multi-Activation model presented in the previous paper, a new Multi-Activation Neural Network has been proposed, which applies multiple activation functions per hidden layer. The model was designed to have the same number of activation functions per layer. As the recently proposed architecture uses five different activation functions, each layer contains a decreasing number of units, a multiple of five [2]. Each layer has different activation functions per unit, so a batch normalization layer is applied after every hidden layer to prevent exploding gradients.

Experiments

This paper is dedicated to testing the performance changes for different ordering of activation functions upon Binary and Multi-Classification datasets. The paper also tests varying Deep Neural Networks to

showcase the effects of a balanced dataset, a dataset with sparse information for training and testing, and another dataset with skewed target values.

The datasets were split into 80 and 20 percent ratios for training and testing, respectively. All the labels in the dataset are normalized using one hot encoder, after which all the numerical values, which are not categorical numerical, are normalized using a standard scalar to homogenize the dataset.

Dataset Description – Heart Failure Prediction

The dataset used for the classification dataset contained 918 observations. The dataset contained varying data types ranging from categorical, categorical numerical, numerical, and labels.

The numerical data attributes consisted of Age, Resting blood pressure, Cholesterol, Maximum heart rate, and Oldpeak, represented as the numerical value for ST. ST is the numerical value measured in depression. The categorical data attributes consisted of Sex, Chest Pain Type, Resting ECG, Exercise Angina, and ST Slope. Although a numerical datatype, according to the data dictionary definition, normal blood sugar is represented as a category. Alternatively, the categorical data type, the target values of if the patient had heart disease or was normal, are represented numerically [2].

The numerical values for Age ranged from 28 to 77, with a median of 53.5 and a standard deviation of 9.43. Although that provides the training models with varying datapoint to prevent overfitting, other features showcase heavy representations for specific categories and ranges of values than others. Within the 918 data observations, 79% of the population was represented by males. The resting blood pressure ranged from 80 mm Hg to 200 mm Hg, with a median of 132 and a standard deviation of 18.5. The highest data variation was present in Cholesterol, as the data ranged from 0 to 603. As the total serum cholesterol contained 19% of the dataset values ranging from 0 to 60, the data for the cholesterol remained untouched [2].

Dataset Description – Cirrhosis Prediction

Cirrhosis showcased the highest range of missing values and data variation in the datasets used for this paper. The missing and NaN values ranged from 6 to 136. Due to missing and nan values, the dataset varied in the count representation of each column. All the rows containing null and NaN values were removed to clean the dataset before using a one-hot encoder and standard scalar normalization. The datasets containing the missing values varied in datatypes from an object, int64, to float64. Using the mean function could resolve the issue with int and float features, but the object features remain unresolved. As most of the nan values were related to the target value 0, the removal of nan values also caused the target values to change. Using the mean function to impute the missing values resulted in repeated data values. Categorical and Label feature columns Status, Drug, Sex, Ascites, Hepatomegaly, Spiders, and Edema were converted to numerical columns. The removal of the nan values also caused the dataset to drop in

data points from 418 total observations, including NaN and missing values, to 276.

Dataset Description – Hepatitis C Prediction

Comparatively to the Cirrhosis dataset, the Hepatitis C dataset contained fewer nan and missing values, ranging from 1 to 18. As the missing values only made up 3% of the dataset, the missing values were deleted from the dataset. The dataset contained two categorial and 11 numerical features.

Although comparatively a cleaner dataset regarding missing values to the Cirrhosis dataset, the target values showcased the dataset as unbalanced. Eighty-nine percent of the target values were “0=Blood Donor”, while “3=Cirrhosis”, “2=Fibrosis”, “1=Hepatitis”, “0s=suspect Blood Donor” made up five, two, three, and one percent respectively. The category feature, the target value for this dataset, was represented as a category/label; they were replaced by values that defined specific categories. “0=Blood Donor”, “0s=suspect Blood Donor”, “1=Hepatitis”, “2=Fibrosis”, and “3-Cirrhosis” were represented as 0, 1, 2, and 3, respectively. As “0s=suspect Blood Donor” represented just a mere one percent of the dataset, it was labeled as 0 as well. The sex categorical feature column was passed through a one-hot encoder to be converted to numerical attributes. The remaining numerical attributes were passed through a standard scaler to normalize the dataset, as each numerical contained various ranges of values and standard deviations.

Test Results and Evaluation

The mentioned datasets were trained and tested using various models, using 500 and 250 epochs, with a batch size of 128. The models used for testing include 5 Uniform, 4 Sequential, and 4 Multi-Activation Deep Neural Networks. The Uniform models contained five hidden layers containing 25, 20, 15, 10, and 5 neurons per layer, respectively, and a SoftMax layer for classification, applying the same activation function throughout the entire model. The 4 Sequential models followed the Breast Cancer Prediction model [5] while moving the sigmoid to a different layer during each variation. The Multi-Activation model employs the recently proposed architecture while changing the location of the activation function Sigmoid during the creation of the model [2].

Over the multiple iterations of training and testing the various models, a common pattern was discovered. The model represented in the paper used for breast cancer prediction [5] would perform the worst consistently throughout the datasets when looking at the training loss.

Heart Failure Dataset

As previously mentioned, the Breast Cancer Detection model, referred to as “Paper MANN” in Figure 1, showcases ill performance during the training phase, while the recently proposed Multi-Activation Neural Network performance nearly as well as a Uniform Tanh model.

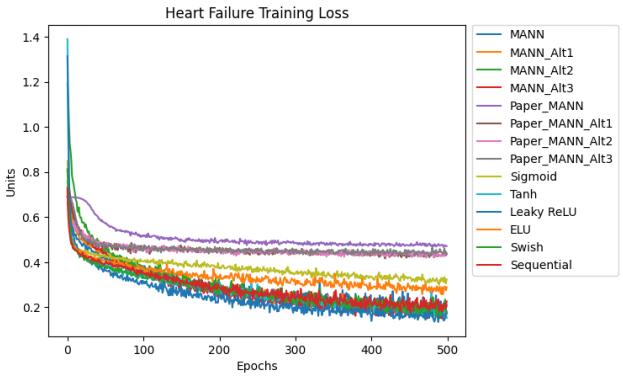


Figure 1 – Heart Failure Classification – Training Loss

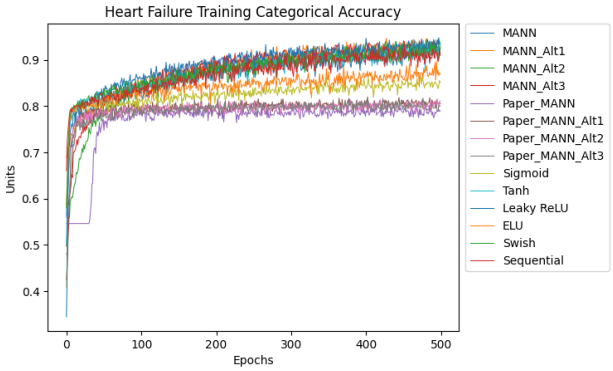


Figure 2 – Heart Failure Classification – Training Accuracy

Looking further into the training accuracy, the Breast Cancer Detection DNN cannot perform as well as the other models. It is possible that the model cannot perform as well as it had done in the paper discussing the model [5] due to the difference in the optimizer used for this model.

Categorical Accuracy	
MANN	74.46%
Paper Alternate MANN 1	74.46%
Paper Alternate MANN 2	73.91%
Sigmoid NN	73.37%
Paper Alternate MANN 3	72.83%
Tanh NN	71.74%
ELU NN	71.20%
Paper MANN	70.65%
Leaky ReLU NN	69.57%
Swish NN	69.57%
Sequential NN	67.39%
MANN_Alt2	66.85%
MANN_Alt3	66.85%
MANN_Alt1	64.13%

Table 1 – Heart Dataset Categorical Accuracy – 500 Epochs

Although the model cannot perform as well as all the other models, the results in Table 1 showcase the effectiveness of activation function ordering. By simply moving the sigmoid function to other hidden layers, it is possible to increase the performance of the models. Merely moving the sigmoid function causes an increase of nearly 4% accuracy in the Paper-based Alternative Sigmoid location 1 model compared to the original model.

Alternatively, it can also be seen how different ordering of the activation functions allows the models to generalize the dataset provided faster than other models.

Cirrhosis Prediction Dataset

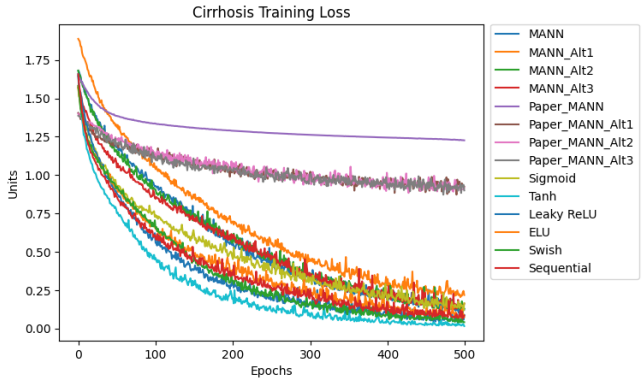


Figure 3 – Cirrhosis Prediction – Training Loss – 500 epochs

As expected, the models cannot learn appropriately due to the sparsity of the data available for training and testing. Although the [5] cannot train as quickly as the Uniform and recently proposed Multi-Activation model, the alternate versions of the "Paper MANN" do exceptionally well compared to other models.

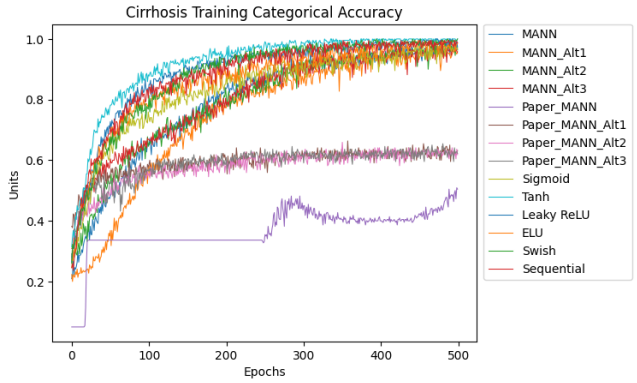


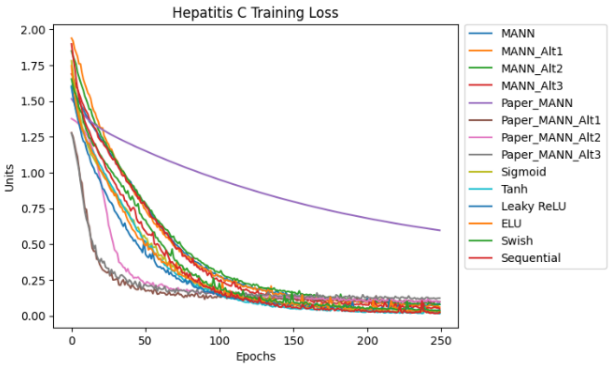
Figure 4 – Cirrhosis Prediction – Training Accuracy – 500 epochs

Categorical Accuracy	
Paper Alternate MANN 1	55.36%
Paper Alternate MANN 3	55.36%
Paper Alternate MANN 2	53.57%
Paper MANN	51.79%
MANN_Alt3	41.07%
Leaky ReLU NN	41.07%
MANN	39.29%
MANN_Alt1	37.50%
Sigmoid NN	37.50%
Tanh NN	37.50%
Sequential NN	33.93%
MANN_Alt2	32.14%
ELU NN	32.14%
Swish NN	26.79%

Table 2 – Cirrhosis Prediction – Testing Accuracy – 500 epochs

As seen from the data in Table 2, changing the ordering of the activation functions can enhance but also provide diminishing results. Comparing the Cirrhosis accuracy to Heart Failure detection during the 500 Epochs training, adjusting the location of the activation functions in the model can provide better results when compared to the base model.

Hepatitis C Prediction Dataset



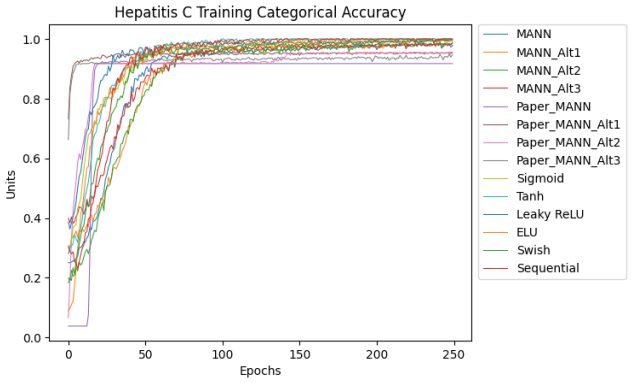


Figure 5 – Hepatitis C Prediction – Training Loss – 500 epochs

Figure 6 – Hepatitis C Prediction – Training Accuracy – 500 epochs

As seen from the training loss and training accuracy, specific altered "Paper MANN" models can adjust to the dataset quickly but perform poorly during the testing phase.

Categorical Accuracy	
Sigmoid NN	94.92%
Sequential NN	94.92%
ELU NN	94.07%
MANN_Alt2	93.22%
MANN_Alt3	92.37%
Paper Alternate MANN 1	92.37%
Leaky ReLU NN	92.37%
Swish NN	92.37%
MANN_Alt1	91.53%
Tanh NN	91.53%
MANN	90.68%
Paper Alternate MANN 2	90.68%
Paper Alternate MANN 3	88.14%
Paper MANN	85.59%

Table 3 – Hepatitis C – Testing Accuracy – 500 epochs

If the "Paper" models are compared, it can be seen that "Paper Alternate MANN 1" has been outperforming the basic "Paper MANN" during every testing phase. It further showcases the effectiveness of the order of Activation functions for the DNNs.

Conclusions

Through multiple iterations of running the models with 500 epochs and a batch size of 128, altering the order of the activation functions can lead to higher accuracy when comparing models with the same architecture. Knowing the benefits of adjusting the order of activation functions, further improvements can be made to the models using the genetic algorithm to test the models dynamically for the best outcome given specific activation functions. Additionally, future testing can be done to analyze and further improve the models by determining which activation functions best fit the dataset using Genetic Algorithms.

Further testing is required to determine the cause of the recently proposed Multi Activation Neural Network Model's performance changing due to the order in which the activation functions are passed to it during creation.

References

[1] Sohal, J. & Stonehill, B. (2023). "Effectiveness of PCA application on a highly complex Dataset" Kennesaw State University

[2] Sohal, J. & Stonehill, B. (2023). "Analysis of Multi-Activation Layers in Neural Network Architectures" Kennesaw State University

[3] Bishop, Chris M., "Neural networks and their applications". *Review of Scientific Instruments* 65. 6(1994): 1803-1832. https://doi.org/10.1063/1.1144830

[4] Anand, R., Mehrotra, K., Mohan, C. K., & Ranka, S. (1995). Efficient classification for multiclass problems using modular neural networks. *IEEE transactions on Neural Networks*, *6*(1), 117-124.

[5] Vijayakumar, K., Kadam, V. J., & Sharma, S. K. (2021). Breast cancer diagnosis using multiple activation deep neural network. *Concurrent Engineering*, *29*(3), 275–284. https://doi.org/10.1177/1063293x211025105