Understanding the Dynamic Sparse Training Models

TEYMURLU IBRAHIM, University of Twente, The Netherlands

Artificial Neural Networks (ANNs) have gained popularity for their improved performance in various fields due to their computational efficiency and reduced storage space requirements. However, traditional ANNs consist of multi-connected layers, leading to an increase in redundant and weak connections as the number of neurons increases. This results in extensive memory and computation consumption. To address this issue, techniques such as pruning and sparsity have been developed. Pruning involves the removal of unnecessary connections that do not have significance in a network, while adaptive/dynamic sparsity involves removing redundant connections while allowing others to grow in their place during training. In this paper, we aim to bring more understanding to Dynamic Sparse Training models by discussing its relation to mammal brains and improving the pruning process and re-connection of neurons. We introduce the Brain-Mimetic Synapse Adjustment algorithm and successfully assess his classification performance using two datasets, CIFAR10 and Fashion MINST. By the end of this research, we expect to contribute to the understanding of Dynamic Sparse Training models, improve the removal process in sparse models, and address the rewiring of neurons in Dynamic Sparse Training models. This gives us valuable insights into developing more human brain-like efficient, and effective neural networks.

Additional Key Words and Phrases: Artificial Neural Network, Brain-Mimetic Synapse Adjustment, Sparse Evolutionary Training, Synaptic Pruning, Optimizing Network Structure

1 INTRODUCTION

Artificial Neural Networks gained popularity for their improved performance in various fields, owing to their ability to keep efficiency in terms of computation power and storage space [28]. Mentioned advantages are giving the ANNs ability to be run on resource-constrained devices [28] which has prohibited resource demands [8]. It has been found that, neural networks closer to mammal brains in structure, are successful in overcoming above stated shortcomings; [8, 16, 22, 28] however, there still exists many limitations regarding the topic.

Human brain produces more than needed neurons when the development process of the brain starts and during childhood synaptic pruning, removes more than half of these synapses [3]. Previous studies show that, brain is removing the weaker connections that are in lesser use [10, 19] to improve the performance [2]. Unlike mammal neural structure, Artificial Neural Networks are, traditionally, consisting of multi-connected layers. That being said, as number of neurons increase, connections increase respectively. Nonetheless, most of these connections are redundant and weak, causing the already pointed out problem of extensive memory and computation consumption [8]. Various techniques have been developed to adress this problem, such as pruning [7, 11, 28] and sparsity [16, 22]. Pruning is removal of unnecessary connections that do not have significance in a network. Sparse neural networks are removing redundant connections while letting others grow in the place [22].

Sparsity can also be seen in a biological brain, where synaptic connections are kept healthy by removal and rewiring [14, 24]. While doing so, the process should be carefully done as errors in the process can cause unwanted results in both human and artificial brain [6, 26]. This paper will stick to sparsity while being inspired by the pruning and addition techniques used in a biological brain.

In this paper, we aim to bring more understanding to Dynamic Sparse Training models by discussing its relation to mammal brains. We are aiming for improvements in pruning process and a better approach in re-connection of the neurons. Having a satisfactory model's performance still remains crucial. Bringing a more systematic approach to determine weakly connected neurons and using these connections for the purpose of rewiring is the most important part of the process. Based on these, our goals can be defined as:

- **Goal 1:** To understand and improve the process of removal of connections with regard to human neural structure, using and perhaps combining previously used approaches [16, 22, 28].
- **Goal 2:** To address the connection of the artificial synapses from a different perspective than already existing, accuracy based connection technique [16].

Following research questions (RQ) will be our main assisting point throughout the research for reaching these goals:

- **RQ1:** How do we need to remove connections in a Dynamic Sparse Traning models to be mimicking biological brain?
- **RQ2:** What are the alternative strategies of adding connections in a Dynamic Sparse Training Model?
- **RQ3:** To what extent can the model performance be improved or kept while following and applying **RQ1** and **RQ2**?

By the end of this research, we contribute in three ways. First and foremost, we bring in more understanding to the Dynamic Sparse Training models. Secondly, we have an improved removal process in sparse models, inspired from the to mammal neural structure. Lastly, we address the rewiring of neurons in Dynamic Sparse Training models, improving the already existing AccSET approach. Furthermore, it is important to point out that keeping the accuracy carries great importance.

The structure of the research paper is as follows. In Section 2, the related work to the topics of pruning in mammalian brains, pruning in neural networks and dynamic sparse neural networks will be further discussed, important points for this research will be pointed out. Section 3 will focus on the methodologies used to address and answer the research questions, followed by a Section 4, which will communicate the setup environment for the experiments. Following this, Section 5 will discuss the results conducted from these experiments, which will be discussed in Section 6.

TScIT 39, July 7, 2023, Enschede, The Netherlands

2 RELATED WORK

In this section, we will point out the related work to the topic of Understanding the Dynamic Sparse Training models. Firstly, previous work on synaptic pruning in mammal brains will be discussed upon. Elaboration on related literature about pruning in artificial and sparse artificial neural networks will follow.

2.1 Synaptic pruning in the human brain

Research conducted on pruning in mammalian brains can be seen in studies earlier than 1998. Chechik et al. [2] approaches this and comes up with a mechanism for synaptic pruning during brain maturation. The research initiates a mathematical model regarding the topic and points out that pruning does not improve performance of the brain. Later he publishes one more paper discussing the removal of weak synapses and modification of the remaining ones [3].

Followingly in 2010, in Herculano et al. [12], he studies connectivity driven white matter scaling and points out that while some of the cerebral cortices show a decrease in connectivity, cerebral cortex may be connected in some other areas.

Later in 2016, it was discussed that microglial processes make connections in synapses after they are being eliminated by Hong et al. [14]. The topic is further discussed in the research, and new insights are found. The important thing to point out in this research for us is the refinement of the synapses.

2.2 Pruning in artificial neural networks

From pruning to compression and later on dynamic sparse training, much work was focused in the last two decades on making the continuously increasing deep learning models more efficient [13].

While focusing on sparse models, there are some distinctive methods worth pointing out. They can be differentiated as:

(1) **Dense-to-sparse**: In this approach, the training starts with a dense neural network, which is then followed by the pruning process. The most popular pruning method is The Lottery Ticket Hypothesis [6].

(2) **Sparse-to-sparse**: In this approach, the training starts with a sparse neural network and is then further trained. This approach introduces two cases:

   (a) *Static sparsity*: A sparse initialization is used, followed by normal training [21].

   (b) *Dynamic sparsity*: A sparse initialization is used, and removals and additions are made during training. Various dynamic sparse training algorithms have been proposed starting from [22]. The main difference between these algorithms lies in how the sparse topology is adapted during training [5, 16, 17].

To point out more on the process of pruning, Han et al. managed to reduce the number of parameters drastically without any loss in accuracy [8]. Later that year, he introduces a new technique called deep compression, where the pruning is one of the three steps. In this case, based on weights, irrelevant connections are found and removed from a network for the purpose of compression [7]. In 2021, Zhao et al. introduces a new dynamical optimization method. He, being inspired by the human brain, keeps track of the unimportant connections for three generations and if the connection keeps performing below a threshold it is being removed [28]. This method makes sure of the irrelevance of a synapse in a network before removing it. Further pruning strategies that are the most popular are the following: magnitude-base, as in SET [22], gradient-base, as in RigL [5], performance-based, as in AccSET [16], a mix of the before-mentioned ones, as in ITOP [17].

Sparsity was another approach used. However, this introduced a new problem in itself, being, what is the optimal level of sparsity and how can it be obtained. Mocanu et al. introduced a Sparse Evolutionary Training algorithm for this purpose in 2018 [22] followed by a PhD research, that was published a year earlier, in which the key concept was dynamic sparsity [20]. SET initially randomly generates the connections, and in each iteration it removes and regrows connections. The removal is done by choosing the values that are closer to zero. In 2020, Lapshyna used SET to improve the regrowing process and came up with a formula for finding the regrow amount while keeping the accuracy [16] and called it AccSET. [1, 23] introduced sampling sparse connectivity based on Bayesian posterior and Dynamic Sparse Representation (DSR) which dynamically adjust the sparsity level respectively.

This paper will mostly refer to SET and AccSET algorithms to understand, analyze and show possible improvements in Dynamic Sparse Training models.

3 METHODOLOGIES

This section is dedicated to explaining the steps that are to be taken for the purpose of answering the above defined research questions. In short, we first improve pruning process based on previously used techniques. Then we refine the way addition of connections are handled. While doing so, we should keep the accuracy of the model.

3.1 Brain-Mimetic Synapse Adjustment

3.1.1 Biological Brain Inspired Pruning. As discussed before, SET and AccSET are removing connections based on their closeness to zero from both, negative and positive sides. However, in a biological brain, this process is a bit different. Synapses in mammal brains are pruned based on the frequency of their usage [25]. Lesser used connections are weakened and then removed while keeping the used connections untouched. In this paper, we apply the biological way of pruning to Sparse Neural Networks.

We start with an already sparse network, which is achieved by generating an Erdős-Rényi sparse weight mask for each layer in the neural network. After each epoch, it is checked which weights are close to zero, as it is an indication of their significance.

The following matrix C_w shown in Algorithm 1 (row 9) is dedicated for saving the insignificance information:

$$C_w = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \\ a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{1}$$

Values noted as a_{mn} where m and n states the index of each weight is in 0 to 3 intervals. Unlike AccSET and SET, the count of appearance

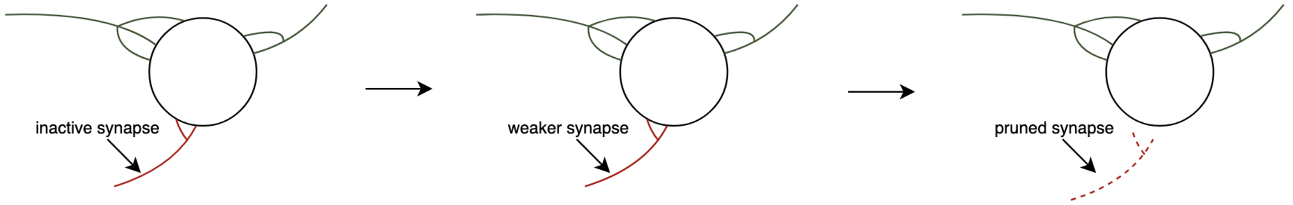


Fig. 1. Illustration of the pruning process in human brain.

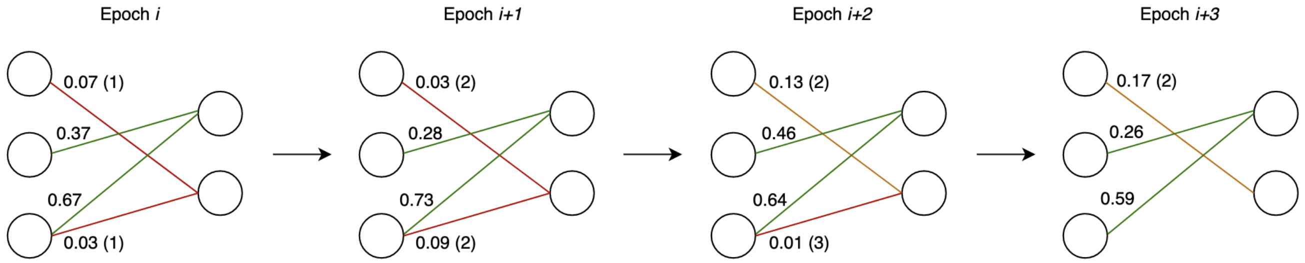


Fig. 2. Illustration of the removal process in an artificial neural network as done in the proposed Brain-Mimetic Synapse Adjustment algorithm..

Algorithm 1 Brain-Mimetic Synapse Adjustment

```
1: Set $\epsilon$ and $\zeta$
2: Initialise ANN model
3: for each fully-connected (FC) layer of the ANN do
4:     Replace FC with a Sparse Connected (SC) layer
5: end for
6: for each training epoch $e$ do
7:     Perform standard training procedure
8:     for each layer do
9:         Pruning candidates that are in $C_w$
10:        if weight appears 3 times then
11:            Restart count for weights under the condition
12:            Remove $\zeta$ weights closest to zero for three generations
13:            $\lambda_e^l$ := current number of connections
14:            $\Delta_e^l \leftarrow \gamma_e^l - \lambda_e^l$
15:            if $e$ is not the last training epoch then
16:                $\theta_e^l \leftarrow 1 - \frac{(acc_l - acc_l \times k)}{k - |acc_l| \times 2 \times k + 1}$
17:                Add $\theta_l^e \times \Delta_l^e$ new connections based on their probability to be added, namely $P$
18:            end if
19:        end if
20:    end for
21: end for
```

of those weights under the condition of being insignificant is kept and if the weights are noted as such for three times consecutively, they are then removed from the network.

Figure 2 and Algorithm 1 (rows 11 and 12) illustrates the process of removal in detail, which has the following steps:

(i) Indicate the connections that have weight value close to zero and mark them (i.e. store them in a list) as shown in *Epoch i* of Figure 2 and Algorithm 1 (row 9)
(ii) Repeat the process while keeping the number of times a connection is insignificant, and check if this number is equal to three.
(iii) If the number is equal to three, it indicates that the connection has been close to zero three times, and therefore we remove the connection as illustrated in *Epoch i+3* of Figure 2.

Following the above stated procedures, just like human neural system, we remove a connection if and only if it has not been used for a while.

3.1.2 Addition of Connections After Being Pruned. After the pruning process takes place, new connections are to be added. AccSET uses accuracy based approach to add connections. Based on the accuracy, number of connections are reduced and added. As accuracy decreases, more connections are added whereas, an increase in accuracy leads to reduction of connections. Our approach is keeping this technique while adding more understanding and logic to it. In section 3.1.1 we have explained how the removal process is improved by keeping track of the insignificant connections. Following this reasoning, we introduce one more matrix, which will be using the before-mentioned one. Each time a connection is removed, we will consider it to be less significant in comparison to one that has been pruned less. Using this, a formula has been developed to improve the addition of artificial neurons in a logical manner.

Unlike most of the dynamic sparse models [SET, rigl, AccSET, . . . etc.], instead of randomly adding connections, we choose which connection to add. Having the significance level of each artificial synapse, we can give each one of them a probability to be added. Less significance indicates a lower probability to be used again, whilst vice versa is the case for connections having more significance level. The formula developed for this purpose follows

$$P = e^{-|k|D} \times (1 + |k|Acc) \tag{2}$$

Here D indicates the matrix that keeps track of the number of times each connection has been removed, and Acc is the current accuracy of the model. Using the Euler's constant, e we are able to exponentially decrease the probability of a connection's addition using $-|k|$ in the power of e. Furthermore, it is important to make sure that more connections are being added when the accuracy is going down. A lower number of neurons in human brain will lead to decreased cognitive ability [12] and therefore, we need to make sure that the addition of artificial synapses is more, especially when the model is performing worse. To assure this, we also include the accuracy in our formula. As the accuracy decreases, the probability of every connection increases. To make sure that probabilities are between 0 and 1 range, we then normalise the probabilities using:

$$N = \frac{P - \min(P)}{\max(P) - \min(P)} = \begin{bmatrix} p_1 & p_2 & \cdots & p_n \\ p_{11} & p_{12} & \cdots & p_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix} \tag{3}$$

The algorithm then calculates the median of the probabilities in N and only adds the connections that are above the median.

4 EXPERIMENTAL SETUP

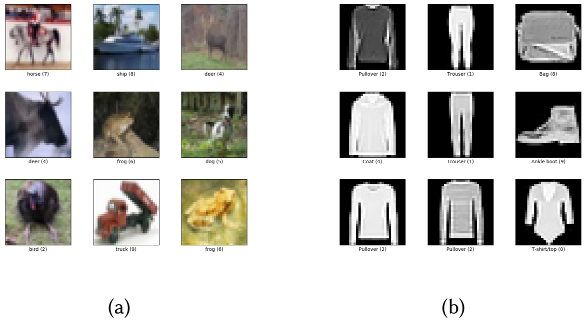
This section will discuss the chosen datasets in detail. That being said, description of the datasets with illustrative images. Baseline approaches, validation metrics will be further explained followed by details of implementation.

4.1 Datasets

The improved and refined algorithms were tested against two datasets, Fashion MNIST [27] and CIFAR10 [15] based on the Multilayer Perceptron (MLP). Section 5 contains the experiments done on mentioned datasets.

4.1.1 Fashion MNIST. Fashion MNIST [27] is an image dataset with 28x28 grayscale images. It consists of 10 classes having 6000 images per one of the following: Tshirts, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots. The examples of images are shown in Figure 3a. It has 60,000 training and 10,000 test samples. We train the model on this dataset for 1000 epochs.

4.1.2 CIFAR10. CIFAR10 [15] dataset just like Fashion MNIST consists of images; however, of datatype RGB colors. It is also consisted of 10 following catagories: airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks which are illustrated in Figure 3b. Train sample size is 50,000 and test sample size is 10,000.



(a)

(b)

Fig. 3. The figure shows example images from both datasets. Figure 3a is samples from CIFAR10 and Figure 3b holds the Fashion MNIST examples.

Overall, although similar in domain these datasets are, they differ in storing the data. Fashion MNIST is easier to train on, considering the low-scale grayscale images and no noise. On the other hand, CIFAR10 is consisting of RGB (i.e 3 color) images with noise and background, which makes it more challenging.

4.2 Baseline Approaches and Validation Metrics

The experiments conducted will be tested against SET [22] and AccSET [16] based on the training on above-discussed datasets. Metrics such as, accuracy, loss, sparsity level (number of connections throughout the training) will be carrying the most importance in terms of validation of effectiveness of the newly proposed algorithm.

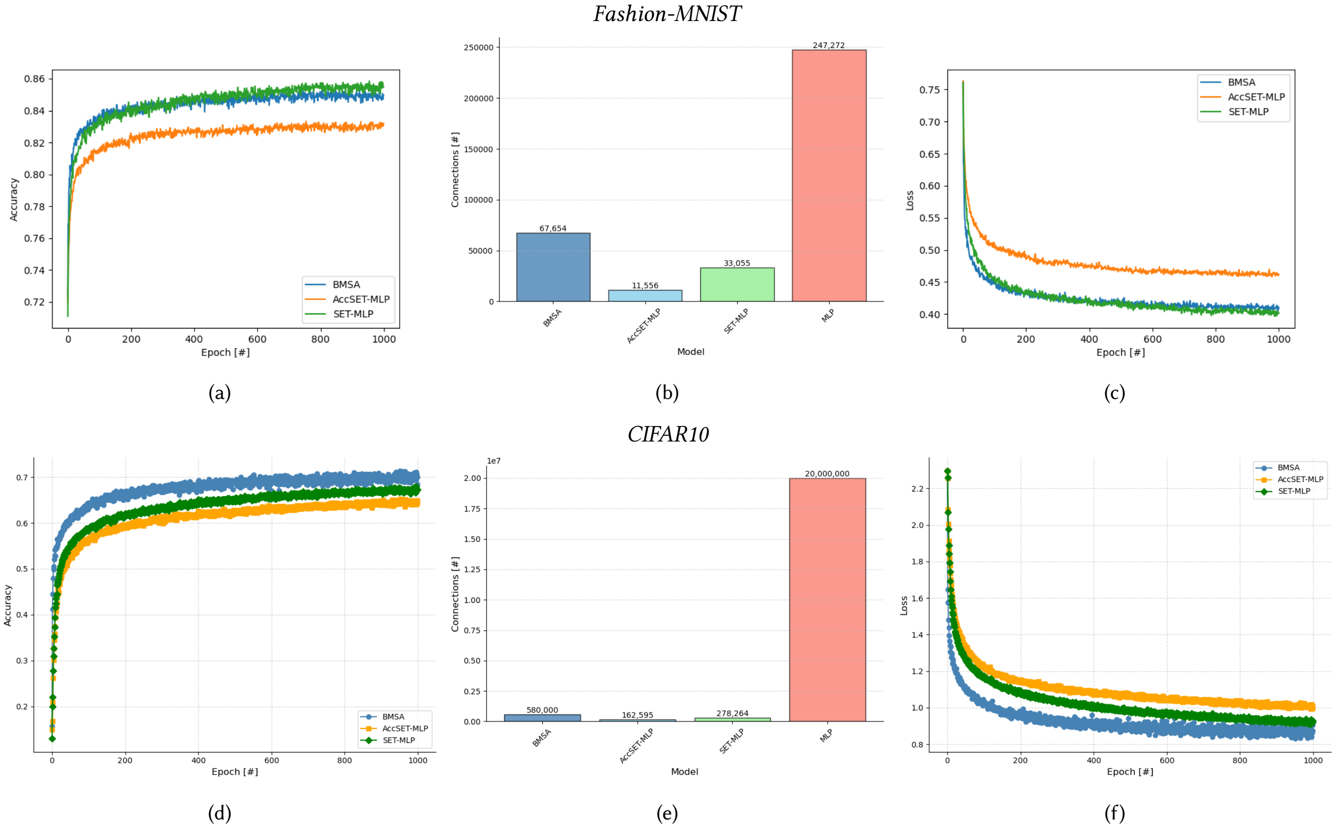
4.3 Implementation Details

It was decided to use Python for the purpose of developing the model. Keras [4] was the chosen implementation method, and the mathematical parts of the model were done using the NumPy library [9].

Table 1. Hyper-parameters used in training of CIFAR10 and Fashion MNIST

Hyper-parameter	Value	Hyper-parameter	Value
Learning rate	0.01	ζ	0.3
Optimiser	SGD	Batch size	100
Momentum	0.9	ϵ	20
Activation Function	LeakyReLU	Loss Function	Categorical Cross-entropy
Dropout rate	0.3		

ζ and ϵ can be seen in both Algorithm 1 and Table 1. ϵ indicates the level of sparsity during initialisation phase of the training process. ζ expresses the sparsity during the training. Both of these variables hold a value between [0, 1] interval. Table 1 show the parameters used. AccSET uses LeakyReLU [18] as an activation model. Better-Prune algorithm is built on top of the AccSET algorithm. The



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 4. The figure illustrate the accuracy of the models (BMSA, SET-MLP, AccSET-MLP) on datasets, Fashion-MNIST 4a and CIFAR10 4d. Comparison in numbers of connections are also shown for both of the datasets 4b, 4e. Lastly, the figure displays the loss of all three models in graphs 4c, 4f.

architercture used for training CIFAR10 dataset is *3072-4000-1000-4000-10* and *784-256-128-100-10* for Fashion MNIST following the settings in [16].

5 RESULTS

Experiments were performed on newly proposed BMSA algorithm and the results were tested against two datasets, Fashion-MNIST and CIFAR10. These factors then were compared against previous algorithms, namely, SET and AccSET to assess the model's performance.

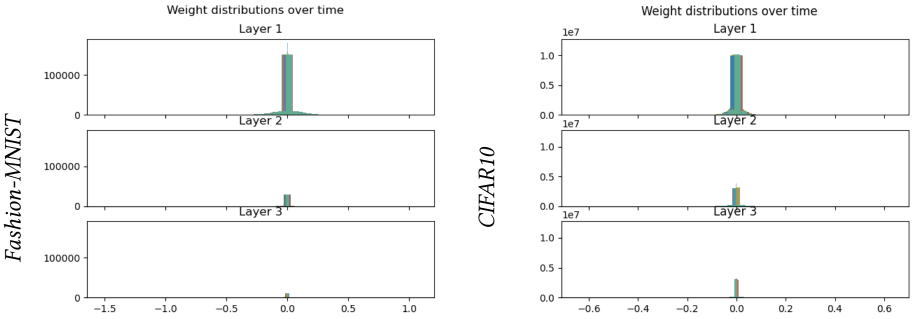
5.1 Fashion-MNIST

Three distinct models, BMSA, ACCSET-MLP and SET-MLP, were compared for accuracy, the result of which can be seen in the graph in Figure 4. The maximum accuracy observed among the models is 0.86 while the minimum is 0.72. During the initial 20 epochs, all three models exhibit a sharp increase in accuracy, which subsequently changes into a fast learning curve. For AccSET-MLP, the stabilization occurs at an accuracy of approximately 0.80 and reaches a maximum of 0.83 by the 1000th epoch. For SET-MLP, the deceleration occurs at an accuracy of approximately 0.82 and continues to increase until reaching 0.86 by the 1000th epoch. BMSA experiences stabilization at an accuracy of approximately 0.82 and remains competitive with SET-MLP until after the 400th epoch when SET-MLP experiences an increase while BMSA remains relatively static at around 0.84, ultimately reaching a maximum of 0.85 by the final epoch. Overall, BMSA shows excellent results, being competitive with SET and out-performing AccSET in terms of accuracy.

Assessment of number of connection for each of the three, BMSA, ACCSET-MLP and SET-MLP models was also performed. The corresponding results are given in Figure 4b below. BMSA has 68000 connections, AccSET-MLP has approximately 11000 connections and SET-MLP has around 35000 connections.

BMSA shows excellent results by having 63% less connections compared to its dense counterpart, MLP.

The results of the relationship between loss and training time (e.g. number of epochs) for the same three models, illustrated in Figure 4c. It is notable from this graph that all the models start of with a loss of approximately 0.75 in the first epoch. By the 10th epoch, BMSA and SET-MLP have both experienced a significant reduction in loss to 0.50 while AccSET-MLP has reduced to 0.58. Subsequently, a fast learning curve is visible for all three models. AccSET-MLP reduces to a loss of 0.49 by the end of the 1000th epoch. BMSA and SET-MLP remain competitive and by the 600th epoch, SET-MLP exhibits a lower loss than BMSA. By the end of the training, the



(a)

(b)

Fig. 5. The figure illustrates the weight distribution of the model BMSA, for each layer. Graph 5a shows the results for Fashion-MNIST and 5b displays the outcomes of CIFAR10

proposed BMSA model shows a better learning behavior compared with SET-MLP.

Furthermore, the weight distribution graph (see Figure 5b) illustrates that the majority of the weights are located at 0, indicating that they have been removed. This observation provides evidence for the sparsity of the model.

5.2 CIFAR10

The accuracy of three models; BMSA, AccSET-MLP, SET-MLP on CIFAR10 dataset are shown in Figure 4d. At the onset, BMSA exhibits superior performance, achieving an accuracy of approximately 0.50. In contrast, AccSET-MLP and SET-MLP only attain this level of accuracy around the 10th epoch. Subsequently, all three models demonstrate a rapid increase in accuracy, with BMSA ultimately reaching a maximum value of 0.70. SET-MLP and AccSET-MLP attain maximum accuracies of 0.67 and 0.63, respectively. Beyond these points, the rate of increase in accuracy diminishes for all models. In summary, BMSA consistently outperforms both AccSET-MLP and SET-MLP in terms of accuracy. Furthermore, BMSA achieves high levels of accuracy more rapidly than the other two models.

The second figure (see Figure 4e), is showing the number of connections in BMSA, AccSET-MLP, SET-MLP and MLP. MLP has the highest number of connections at 2,000,000, while BMSA has significantly fewer connections with a total of 580,000. SET-MLP and AccSET-MLP have even fewer connections, with totals of 278,264 and 162,595, respectively. Comparing BMSA and MLP in terms of the percentage of connections, BMSA has only 20.4% of the entire connections. Overall, BMS use 5 times less number of connections compared to their dense counterpart.

The Figure 4f, illustrates the loss per epoch for three models. Initially, BMSA has a lower loss of around 1.2, while the other two models start with a loss of around 1.3. Over time, the loss for all three models gradually decreases. By the 1000th epoch, BMSA has the lowest loss of around 0.87, while SET-MLP and AccSET-MLP have losses of around 0.92 and 1.01, respectively. Overall, it can be observed that BMSA consistently has the lowest loss, followed by SET-MLP and then AccSET-MLP. This indicates that BMSA is the most effective model in terms of minimizing loss.

Just like the Fashion-MNIST dataset, the sparsity level stays high on CIFAR10 as well. This can be seen from the Figure 5b where most of the weights are dense around 0. Be reminded that, 0 indicates no connection.

6 DISCUSSIONS

The comparison of the models, BMSA, ACCSET-MLP and SET-MLP, enabled us to validate the possibility of removing and adding connections in a Dynamic Sparse Training model by mimicking biological brain. The study also verified whether the proposed BMSA algorithm can improve performance of Dynamic Sparse Training model.

6.1 Insights on Brain-Mimetic Pruning

The present study introduces a novel approach called Brain-Mimetic Pruning, which incorporates principles from the biological brain's synaptic pruning process into the context of Sparse Neural Networks. Unlike traditional pruning methods such as AccSET and SET that remove connections based on their proximity to zero, this approach emulates the biological brain's mechanism of pruning synapses based on their frequency of usage.

The pruning process involves identifying connections with weight values close to zero and marking them for potential removal. These connections are tracked, and if they remain insignificant for three consecutive times, they are removed from the network. This mimics the biological brain's behavior of pruning connections that have not been frequently utilized, allowing the network to adapt its connectivity based on usage patterns.

By incorporating this biological pruning mechanism, the study aims to enhance the efficiency and performance of Sparse Neural Networks. By removing unused connections, the network can allocate its computational resources more effectively and potentially reduce overfitting. Additionally, the sparsity induced by this pruning method contributes to the interpretability of the network, as the majority of connections carry meaningful information.

The successful implementation of Brain-Mimetic Pruning holds several implications for the field of deep learning. Firstly, it offers a biologically inspired alternative to traditional pruning methods, allowing for more efficient utilization of computational resources. Secondly, the approach introduces a level of interpretability to Sparse Neural Networks by emphasizing the importance of frequently used connections. This could aid in understanding the network's decision-making process and facilitate model explanation.

While this study demonstrates the feasibility and potential benefits of Brain-Mimetic Pruning, further research is warranted to explore its applicability in different domains and datasets.

6.2 Insights on Brain-Mimetic Addition

In this study, we also address the issue of adding connections after the pruning process in Sparse Neural Networks. While traditional approaches like AccSET use accuracy-based methods to determine the number of connections to add or remove, we aim to enhance this technique by introducing a more logical and nuanced approach.

Unlike many dynamic sparse models that randomly add connections, our approach focuses on selecting which connections to add. With the knowledge of each artificial synapse's significance level, we assign a probability to each connection for potential addition. Connections with lower significance are assigned a lower probability of being used again, while connections with higher significance have a higher probability of being added.

Furthermore, to strike a balance between adding too few or too many connections, the algorithm calculates the median of the probabilities. Only the connections with probabilities above the median are selected for addition. This approach ensures a controlled and balanced increase in the number of connections, taking into account both the significance of connections and the current accuracy of the model.

The successful implementation of this approach holds several implications for the field of deep learning. By adding connections in a more strategic and informed manner, the network can adapt its architecture to better capture complex patterns and improve its overall performance. Moreover, the use of a logical framework for connection addition contributes to the interpretability of Sparse Neural Networks, as connections with higher significance carry more weight in the decision-making process.

6.3 Insights on Model Performance

The performance of BMSA was compared to two distinct models, namely AccSET-MLP, and SET-MLP, on two widely-used image classification datasets, Fashion-MNIST and CIFAR10. The evaluation was based on accuracy, number of connections, and loss metrics, providing valuable insights into the strengths and limitations of each BMSA.

The results of accuracy demonstrate that BMSA is competitive with SET-MLP in terms of accuracy. Both models experience a rapid increase in accuracy during the initial 20 epochs and remain competitive until after the 400th epoch, when SET-MLP experiences an increase while BMSA remains relatively static. By the final epoch, BMSA achieves a maximum accuracy of 0.85 which is considerably higher than AccSET and only slightly lower than the maximum accuracy achieved by SET-MLP (0.86). On the CIFAR10 dataset, BMSA exhibited superior performance by achieving a maximum accuracy of 0.70, surpassing the accuracies of SET-MLP (0.67) and AccSET-MLP (0.63). These results indicate that BMSA has a higher potential for accurate image classification compared to the other models.

The number of connections analysis revealed interesting differences in the model architectures. BMSA exhibited the highest number of connections among the three models, indicating a higher level of complexity and connectivity in its architecture. It can be seen that, BMSA has significantly fewer connections compared to its dense counterpart, MLP. BMSA has a more intricate model structure and potentially has the ability to capture more intricate patterns and relationships within the datasets. Our results show that, in a more complex dataset, CIFAR10, BMSA outperformed its counterparts while keeping up with SET against Fashion-MNIST.

Furthermore, the analysis of loss values provided insights into the models' optimization capabilities. BMSA consistently achieved lower loss values, indicating its effectiveness in minimizing errors during the training process. Specifically, BMSA demonstrated the lowest loss values on both Fashion-MNIST (final loss of 0.42) and CIFAR10 (final loss of 0.87), followed closely by SET-MLP. AccSET-MLP exhibited slightly higher losses on both datasets. These findings highlight the superior optimization abilities of BMSA in terms of reducing the discrepancy between predicted and actual labels.

In conclusion, the comparative analysis suggests that BMSA outperforms AccSET-MLP and SET-MLP in terms of accuracy, loss minimization, and convergence speed on both Fashion-MNIST and CIFAR10 datasets. BMSA consistently achieved higher accuracies, demonstrated superior optimization capabilities by minimizing loss, and showcased faster convergence to high accuracy levels. These findings have significant implications for the field of image classification, as BMSA can be considered a more effective and efficient model for accurate classification tasks. However, further research is needed to investigate the generalization capabilities and potential trade-offs associated with the increased complexity of BMSA's architecture.

7 CONCLUSION

In this research, we presented a novel approach for the pruning and growth processes in Sparse Neural Networks, inspired by the biological brain. Our proposed Brain-Mimetic Synapse Adjustment (BMSA) algorithm incorporates. the concept of significance and frequency of usage to prune connections and add new ones, mimicking the pruning mechanisms observed in mammalian brains. The findings of this research have significant implications for the field of deep learning. Our approach provides a more biologically plausible and intuitive method for network pruning and growth. The combination of improved performance, interpretability, and biological inspiration makes our approach a promising direction for future research in sparse network modeling.

Despite the success of our Biological Brain Inspired Pruning approach, there are limitations to consider. Time constraints hindered experimentation and exploration of alternative methods for determining connection insignificance.

Further work may include: (1) exploration of the generalization of BMSA learning capabilities under various datasets, (2) investigation of other biological-inspired possibile algorithmic improvements of

BMSA, and (3) fast adaptation of BMSA algorithm to other dynamic sparse training models (e.g. RigL).

REFERENCES

[1] Guillaume Bellec, David Kappel, Wolfgang Maass, and Robert Legenstein. 2018. Deep Rewiring: Training very sparse deep networks. arXiv:1711.05136 [cs.NE]

[2] Gal Chechik, Isaac Meilijson, and Eytan Ruppin. 1998. Synaptic Pruning in Development: A Computational Account. *Neural Computation* 10, 7 (1998), 1759–1777. Publisher: MIT Press.

[3] Gal Chechik, Isaac Meilijson, and Eytan Ruppin. 1999. Neuronal Regulation: A Mechanism for Synaptic Pruning During Brain Maturation. *Neural Computation* 11, 8 (Nov. 1999), 2061–2080. Publisher: MIT Press.

[4] François Chollet and others. 2015. Keras.

[5] Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. 2020. Rigging the Lottery: Making All Tickets Winners. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2943–2952. ISSN: 2640-3498.

[6] Jonathan Frankle and Michael Carbin. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. arXiv:1803.03635 [cs].

[7] Song Han, Huizi Mao, and William J. Dally. 2016. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. arXiv:1510.00149 [cs].

[8] Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both Weights and Connections for Efficient Neural Network. In *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc.

[9] Charles R. Harris, K. Jarrod Millman, St\éfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern\ández del R\ío, Mark Wiebe, Pearu Peterson, Pierre G\érard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362. Publisher: Springer Science and Business Media LLC.

[10] Akiko Hayashi-Takagi, Sho Yagishita, Mayumi Nakamura, Fukutoshi Shirai, Yi I. Wu, Amanda L. Loshbaugh, Brian Kuhlman, Klaus M. Hahn, and Haruo Kasai. 2015. Labelling and optical erasure of synaptic memory traces in the motor cortex. *Nature* 525, 7569 (Sept. 2015), 333–338. Number: 7569 Publisher: Nature Publishing Group.

[11] Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. 2019. Filter Pruning via Geometric Median for Deep Convolutional Neural Networks Acceleration. arXiv:1811.00250 [cs].

[12] Suzana Herculano-Houzel, Bruno Mota, Peiyan Wong, and Jon H. Kaas. 2010. Connectivity-driven white matter scaling and folding in primate cerebral cortex. *Proceedings of the National Academy of Sciences* 107, 44 (Nov. 2010), 19008–19013. Publisher: Proceedings of the National Academy of Sciences.

[13] Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. 2021. Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research* 22, 241 (2021), 1–124.

[14] Soyon Hong, Lasse Dissing-Olesen, and Beth Stevens. 2016. New insights on the role of microglia in synaptic pruning in health and disease. *Current Opinion in Neurobiology* 36 (Feb. 2016), 128–134.

[15] A. Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images.

[16] V. Lapshyna. 2020. Sparse artificial neural networks : Adaptive performance-based connectivity inspired by human-brain processes. Publisher: University of Twente.

[17] Shiwei Liu, Lu Yin, Decebal Constantin Mocanu, and Mykola Pechenizkiy. 2021. Do We Actually Need Dense Over-Parameterization? In-Time Over-Parameterization in Sparse Training. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 6989–7000. ISSN: 2640-3498.

[18] Andrew L. Maas. 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models.

[19] Constantine A. Mangina and Evgeni N. Sokolov. 2006. Neuronal plasticity in memory and learning abilities: Theoretical position and selective review. *International Journal of Psychophysiology* 60, 3 (June 2006), 203–214.

[20] Decebal Constantin Mocanu. 2017. Network computations in artificial intelligence. (June 2017).

[21] Decebal Constantin Mocanu, Elena Mocanu, Phuong H. Nguyen, Madeleine Gibescu, and Antonio Liotta. 2016. A topological insight into restricted Boltzmann machines. *Machine Learning* 104, 2 (Sept. 2016), 243–270.

[22] Decebal Constantin Mocanu, Elena Mocanu, Peter Stone, Phuong H. Nguyen, Madeleine Gibescu, and Antonio Liotta. 2018. Scalable Training of Artificial Neural Networks with Adaptive Sparse Connectivity inspired by Network Science. *Nature Communications* 9, 1 (June 2018), 2383. arXiv:1707.04780 [cs].

[23] Hesham Mostafa and Xin Wang. 2019. Parameter Efficient Training of Deep Convolutional Neural Networks by Dynamic Sparse Reparameterization.

[24] Rosa C. Paolicelli, Giulia Bolasco, Francesca Pagani, Laura Maggi, Maria Scianni, Patrizia Panzanelli, Maurizio Giustetto, Tiago Alves Ferreira, Eva Guiducci, Laura Dumas, Davide Ragozzino, and Cornelius T. Gross. 2011. Synaptic Pruning by Microglia Is Necessary for Normal Brain Development. *Science* 333, 6048 (Sept. 2011), 1456–1458. Publisher: American Association for the Advancement of Science.

[25] Joshua R. Sanes and Jeff W. Lichtman. 1999. Development of the Vertebrate Neuromuscular Junction. *Annual Review of Neuroscience* 22, 1 (1999), 389–442. _eprint: https://doi.org/10.1146/annurev.neuro.22.1.389.

[26] Carl M. Sellgren, Jessica Gracias, Bradley Watmuff, Jonathan D. Biag, Jessica M. Thanos, Paul B. Whittredge, Ting Fu, Kathleen Worringer, Hannah E. Brown, Jennifer Wang, Ajamete Kaykas, Rakesh Karmacharya, Carleton P. Goold, Steven D. Sheridan, and Roy H. Perlis. 2019. Increased synapse elimination by microglia in schizophrenia patient-derived models of synaptic pruning. *Nature Neuroscience* 22, 3 (March 2019), 374–385.

[27] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms.

[28] Feifei Zhao and Yi Zeng. 2021. Dynamically Optimizing Network Structure Based on Synaptic Pruning in the Brain. *Frontiers in Systems Neuroscience* 15 (2021), 620558.