
Evolving and Understanding Sparse Deep Neural Networks using Cosine Similarity

Joost Pieterse¹ Decebal Constantin Mocanu¹

Abstract

Training sparse neural networks with adaptive connectivity is an active research topic. Such networks require less storage and have lower computational complexity compared to their dense counterparts. The Sparse Evolutionary Training (SET) procedure uses weights magnitude to evolve efficiently the topology of a sparse network to fit the dataset, while enabling it to have quadratically less parameters than its dense counterpart. To this end, we propose a novel approach that evolves a sparse network topology based on the behavior of neurons in the network. More exactly, the cosine similarities between the activations of any two neurons are used to determine which connections are added to or removed from the network. By integrating our approach within the SET procedure, we propose 5 new algorithms to train sparse neural networks. We argue that our approach has low additional computational complexity and we draw a parallel to Hebbian learning. Experiments are performed on 8 datasets taken from various domains to demonstrate the general applicability of our approach. Even without optimizing hyperparameters for specific datasets, the experiments show that our proposed training algorithms usually outperform SET and state-of-the-art dense neural network techniques. The last but not the least, we show that the evolved connectivity patterns of the input neurons reflect their impact on the classification task.

1. Introduction

Dense artificial neural networks are a commonly used machine-learning technique in deep learning that has a wide range of applications. Yet, they have multiple limitations,

¹Department of Mathematics and Computer Science, Eindhoven University of Technology, Netherlands. Correspondence to: Joost Pieterse <joost.pieterse@outlook.com>.

Preprint version. Code available online:
<https://github.com/joostPieterse/CosineSET>

several of which are potentially addressable by sparse artificial neural networks. Most existing research on this topic focuses on reducing storage and prediction time, for example to be able to use neural networks in embedded devices (Han et al., 2015; Srinivas et al., 2017). We, however, are interested in algorithms that reduce training time as well. By reducing memory requirements and training time, usage of neural networks is made more accessible. It could for example facilitate training neural networks in those embedded devices. Furthermore, it may allow for deploying very large networks, such that they can be used to tackle datasets with a large number of features directly.

This objective makes approaches such as (Han et al., 2015; Srinivas et al., 2017) unsuitable for our purposes, as they still use the full dense network during training. There are also approaches that do not use the dense network, but instead depend on defining the network's topology before the training phase (Dey et al., 2017; Mocanu et al., 2016). This pre-defined sparsity may however not be optimal for all datasets. For this reason, we consider an alternative approach that does not use the full network and that does not rely on a pre-defined network topology: Sparse Evolutionary Training (SET) (Mocanu et al., 2018). SET starts out with a randomly generated sparse network and updates its topology after each training epoch based on the values of its weights. In SET's experiments, training a network using SET gave better results than its nonevolutionary (i.e. using pre-defined sparsity) counterpart, as the final topology is better suited to the training data. Additionally, its results were most of the times even better than those of its densely connected counterpart. In the original paper, it was already suggested that the algorithm may still be improved by using alternative techniques to evolve the network, such as preferential attachment.

In this paper, we take a novel direction and propose to evolve the network's topology by using domain knowledge. This is also our main contribution: to determine the importance of a connection based on the cosine similarity between the activations of the two neurons of that connection. Reason for this is that if the cosine similarity is close to zero, this is an indication that there is no meaningful relation between these neurons. We then propose systematically five new

algorithms that use this technique to replace the original procedures for adding and removing connections in SET. On top of that, we analyze the additional computational complexity of our method and argue that it should not cause noticeable overhead, while suggesting methods to further reduce complexity for extreme cases. Next, each algorithm is tested on 8 different datasets in order to demonstrate the improvement of our approach over SET. Our results show that usually our algorithms outperform SET and dense state-of-the-art neural network techniques, while having many times less parameters than the latter ones. They also reveal that using cosine similarity to evolve a network may reduce overfitting. Finally, we show the evolved connectivity patterns of the input neurons (or input features) reflect very well their impact on the classification task and may contribute to further understanding the behavior of neural networks with adaptive sparse connectivity.

2. Background

There are many types of neural networks, such as Multi-Layer Perceptrons (MLPs), Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) (LeCun et al., 2015). In this paper we will focus on the most vanilla type of neural networks, MLPs, as they represent 61% of a typical Google TPU (Tensor Processing Unit) workload for production neural networks applications, while convolutional neural networks represent just 5% (Jouppi et al., 2017). A dense MLP consists of a number of fully connected bipartite layers. In contrast to these commonly used dense networks, there is also research into sparsely connected networks. Dense neural networks have been shown to have a large number of redundant parameters, in some cases more than 95% of the parameters can be predicted from the remaining ones without accuracy loss (Denil et al., 2013). In early work on sparsification, Optimal Brain Damage (LeCun et al., 1990) and Optimal Brain Surgeon (Hasibi et al., 1993) use gradient methods in order to sparsify networks during training. They noted that a sparse network has several advantages over its dense counterpart, such as better generalization, reduced memory footprint and improved prediction time.

More recently, (Han et al., 2015) proposed a magnitude-based method for obtaining a sparse network. After pruning the dense network during training, the network is retrained in order to improve accuracy. Their motivation for employing magnitude-based pruning is that alternatives such as using second order derivatives are computationally intensive. In (Srinivas et al., 2017), gate variables are introduced that represent whether a connection is present. These gate variables are parameters that are optimized during training, and as such introduce additional overhead.

Note that all of these approaches do use the full dense net-

work during training. In (Dey et al., 2017), a method for obtaining a sparse neural network was introduced, which does not require training the dense network. Based on the user-specified number of connections per neuron, the topology is determined by an interleaver algorithm ensuring good spatial spread of connections. Although this pre-determined sparsity may allow for larger networks, it is not flexible for handling a wide range of datasets with various characteristics.

In (Mocanu, 2017; Mocanu et al., 2018) SET is introduced, while variants of it are discussed for federated learning in (Zhu and Jin, 2018) and image classification in (Mostafa and Wang, 2019). SET is an algorithm for training sparse neural networks with adaptive connectivity. Like the previously described approaches, SET starts out with a sparsely connected network. The topology of this network, however, is not static but instead evolves during training. After each training epoch, when the weights have been trained to a reasonable level to suit the provided data, the connections having weights closest to zero are removed (weights magnitude based removal). New connections replacing removed connections are randomly selected and added to the network. As the evolution of the network is specific to the data, this approach is more flexible. This was also revealed in their results, in which SET outperforms both dense networks and static (i.e. nonevolutionary) sparse networks. The original SET algorithm uses a straightforward randomized method for evolving the topology of a neural network, while encouraging research into more sophisticated methods. The interested reader is referred to (Mocanu et al., 2018) for more details on SET. Further on, we provide background on cosine similarity (Tan et al., 2006), a similarity measure that is used in our method. Cosine similarity is defined as the cosine of the angle between two vectors. It has various applications, such as text classification, document clustering and face verification. An important reason for adopting cosine similarity is the fact that it is efficient to evaluate. Another important property is that the length of each vector that is being compared is normalized.

3. Proposed methods

This section details our proposed approach. First, it presents how cosine similarity can be used to determine the importance of neural network connections. Second, it introduces 5 new algorithms that integrate cosine similarity based connections importance into the SET procedure. To the end, it discusses the computational complexity and the relation to Hebbian learning of our approach.

3.1. Cosine similarity to detect connections importance

The basic idea is that the sparse network topology can be evolved based on the behavior of its neurons. The impor-

tance of a connection is determined by the similarity of the activations of these neurons, which were obtained during the feedforward phase of an epoch. The similarity measure that we employ is cosine similarity. For activation vectors \mathbf{a} and \mathbf{b} , this is defined as:

$$\frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \quad (1)$$

Intuitively, if two neurons exhibit similar behavior, this indicates that the value of one neuron can help predict the value of the other neuron and can therefore aid in propagating patterns present in the data. For this reason, this connection can help to establish the behavior of the receiving neuron in the simplest way possible. Thus, a connection is more likely to be meaningful if there is a consistent relation between the behavior of the two neurons. We argue that cosine similarity is a suitable way of determining if such a relation exists, since it can consider two vectors to be similar when the signs of the activations frequently agree, but the magnitude does not. This is desirable as a consistent difference in magnitude can be mitigated by the weight of a connection. We will further introduce systematically five new algorithms to evolve sparse neural networks using cosine similarity. For consistency, the notation introduced in (Mocanu et al., 2018) is used. Let n^k denote the number of neurons of layer k in the neural network with s available training samples. ε is a constant controlling the sparsity of the network. Each bipartite layer $n^{k-1} \times n^k$ has weight matrix $\mathbf{W}^k \in \mathbf{R}^{n^{k-1} \times n^k}$.

3.2. Proposed algorithms

3.2.1. COSINE SIMILARITY-BASED CONNECTION ADDITION (CoDASET)

The neural network is initialized in the same way as the network of SET: each connection in a bipartite layer with $n^{k-1} \times n^k$ neurons exists with probability $\frac{\varepsilon(n^{k-1} + n^k)}{n^{k-1}n^k}$, resulting in a sparse network. In the training phase, the weights are updated by stochastic gradient descent. However, we also fill out an activation matrix $\mathbf{A}^k \in \mathbf{R}^{n^k \times s}$ for each layer k during the feedforward phase. After each training epoch, the cosine similarity matrix $\mathbf{C}^k \in \mathbf{R}^{n^{k-1} \times n^k}$ can then be calculated as follows:

$$\mathbf{C}_{pq}^k = \left| \frac{\mathbf{A}_p^{k-1} \cdot \mathbf{A}_q^k}{\|\mathbf{A}_p^{k-1}\| \|\mathbf{A}_q^k\|} \right| \quad (2)$$

where p and q are neurons in layer $k-1$ and k respectively and \mathbf{A}_p^k is the activation vector for neuron p of length s . This is followed by the rewiring step, in which the connections having weight closest to zero are removed. Since \mathbf{C}^k contains absolute values, we can retrieve the set of connections with highest similarity from this matrix and add these con-

nections to the network. Pseudocode for this approach can be found in Algorithm 1.

Algorithm 1 CoDASET pseudocode

```

initialize SET model
for each training epoch  $e$  do
    perform standard feedforward phase, storing activations in activation matrix  $\mathbf{A}$ 
    backpropagate and perform weights update
    for each bipartite SC layer  $k$  do
        remove a fraction  $\zeta$  of the smallest positive weights
        remove a fraction  $\zeta$  of the highest negative weights
        calculate cosine similarity matrix  $\mathbf{C}$  according to equation 2
        add connections with largest value in  $\mathbf{C}$  in the same amount as previously removed
    end for
end for
    
```

3.2.2. COSINE SIMILARITY-BASED PROBABILISTIC CONNECTION ADDITION (CoPASET)

Here, we propose a probabilistic variant of CoDASET. New connections are not chosen by using cosine similarity directly, but instead by drawing from a probability distribution based on the cosine similarities of the connections. Each connection has a probability of being added to the network that corresponds to its normalized cosine similarity:

$$P(\mathbf{W}_{pq}^k) = \frac{\mathbf{C}_{pq}^k}{\sum_{i=0}^{n^{k-1}} \sum_{j=0}^{n^k} \mathbf{C}_{ij}^k} \quad (3)$$

where $P(\mathbf{W}_{pq}^k)$ is the probability of adding a connection between neurons p and q in bipartite layer k . This method reintroduces randomness into the connection selection procedure and may therefore lead to a better exploration of possible topologies, as CoDASET can potentially get stuck in a local minima (adding and removing the same connections after each epoch). On the other hand, we do expect to select more interesting connections than SET by using a probability distribution proportional to cosine similarity. Pseudocode is presented in Algorithm 2.

3.2.3. COSINE SIMILARITY-BASED CONNECTION REMOVAL (CoRSET)

SET removes connections based on their weight. For such magnitude-based methods it has been shown that they often remove the wrong connections, e.g. by Optimal Brain Surgeon (Hassibi et al., 1993). Unfortunately, their proposed alternative is computationally expensive. Therefore, we also

Algorithm 2 CoPASET pseudocode

```

initialize SET model
for each training epoch  $e$  do
    perform standard feedforward phase, storing activations in activation matrix  $\mathbf{A}$ 
    backpropagate and perform weights update
    for each bipartite SC layer  $k$  do
        remove a fraction  $\zeta$  of the smallest positive weights
        remove a fraction  $\zeta$  of the highest negative weights
        calculate cosine similarity matrix  $\mathbf{C}$  according to equation 2
        create probability distribution following equation 3
        add connections by drawing samples in the same amount as previously removed
    end for
end for
    
```

apply cosine similarity to connection elimination. Instead of eliminating the connections that have smallest weight, the connections for which the product of their weight and cosine similarity is smallest are eliminated:

$$\mathbf{M}_{pq}^k = \mathbf{W}_{pq}^k \mathbf{C}_{pq}^k \quad (4)$$

where pq is an existing connection in layer k . Note that this method does not require computing \mathbf{C}^k fully, since only the values for existing connections are used. We hypothesize that using cosine similarity as an additional indicator for the importance of a connection can result in improved connection removal. Algorithm 3 shows pseudocode for this method.

3.2.4. CoDACoRSET

This method combines CoDASET with CoRSET, i.e. it uses CoRSET to remove the unimportant connections, and CoDASET to add new connections.

3.2.5. COPACoRSET

This method combines CoPASET with CoRSET, i.e. it uses CoRSET to remove the unimportant connections, and CoPASET to add new connections.

3.3. Computational Complexity

An important property of SET is the potential to reduce both training and prediction time by capitalizing on the sparsity of the network. For this reason, we analyze the additional overhead which our proposed approach brings, such that a trade-off can be made between computational complexity and any possible accuracy improvements. The extra overhead mainly resides in calculating the activation matrix \mathbf{C} . For each entry in this matrix, 3 dot products between vectors of size s are performed. CoRSET, however, does not require calculating the full matrix, since only the cosine similarities of existing connections are needed. Therefore, it requires $3s \frac{\varepsilon(n^{k-1} + n^k)}{n^{k-1}n^k}$ computations per bipartite layer $n^{k-1} \times n^k$. CoDASET and CoPASET on the other hand do

use the full matrix, thus those need $3sn^{k-1}n^k$ computations per bipartite layer $n^{k-1} \times n^k$. This is comparable to a single feedforward phase of a dense network, in which such a bipartite layer calculates $sn^{k-1}n^k$ products. Since the back-propagation phase dominates the feedforward phase, we argue that our method should not cause noticeable overhead in an efficient implementation.

3.4. Relation to Hebbian Learning

Hebbian learning (Hebb, 1949) is an alternative learning algorithm for neural networks. Instead of updating weights using global information from a loss function, weights are updated solely based on the values of the activations of the pre- and post-synaptic neuron of that connection, i.e. in a local manner. Pre-synaptic neurons refer to neurons sending their activations over a connection in a bipartite layer, with post-synaptic neurons being the receivers. Using this terminology, the principle that Hebbian learning is based on can be stated as the theory that a post-synaptic neuron is more responsive to activations of pre-synaptic neurons that frequently take part in firing this neuron. In other words, a connection is strengthened if the activations of its neurons agree, this is often summarized as "fire together, wire together". Following this principle, weights are updated by Hebb's rule:

$$\dot{w}_{ij} = \eta x_i y_j \quad (5)$$

where \dot{w}_{ij} is the change in weight magnitude, η is the learning rate, and x_i and y_j the activations of neurons i and j respectively. As (Wadhwa and Madhow, 2016) pointed out, Hebbian learning is largely ignored for machine learning tasks. Yet, it has some interesting properties, such as being biologically plausible. For this reason, we discuss the similarities: the cosine similarity of a connection in our method is also determined by multiplying the activations of its neurons, albeit normalized. So if the activations of two connected neurons agree, both Hebbian learning and our methods would increase the importance of this connection, respectively resulting in increased weight and a better chance at adding or preserving this connection. Thus, both methods reward connections between neurons that exhibit similar behavior. There is, however, a difference in usage: Hebb's rule is used to optimize the network's parameters values, whereas cosine similarity here is employed in order to evolve the network's topology.

4. Experiments

4.1. Setup

Experiments were performed on several datasets retrieved from the UCI Machine Learning Repository¹: MicroMass (Mahé et al., 2014), CNAE-9 (Ciarelli and Elias, 2012),

Algorithm 3 CoRSET pseudocode

```

initialize SET model
for each training epoch e do
    perform standard feedforward phase, storing activations in activation matrix A
    backpropagate and perform weights update
    for each bipartite SC layer k do
        for each existing connection pq in k do
             $\mathbf{C}_{pq}^k \leftarrow \frac{\mathbf{A}_p^{k-1} \cdot \mathbf{A}_q^k}{|\mathbf{A}_p^{k-1}| |\mathbf{A}_q^k|}$ 
        end for
        %calculate metric for removal
        for each existing connection pq in k do
             $\mathbf{M}_{pq}^k \leftarrow \mathbf{W}_{pq}^k \mathbf{C}_{pq}^k$ 
        end for
        remove fraction  $\zeta$  of connections from the network with smallest metric value in M
        randomly add connections in the same amount as previously removed
    end for
end for
    
```

¹<http://archive.ics.uci.edu/ml/>

Epilepsy (Andrzejak et al., 2001), Human Activity Recognition (HAR) (Anguita et al., 2013), Madelon (Guyon et al., 2005) and ISOLET (Cole, 1990). Additionally, two image datasets were used: COIL-100 (Nayar et al., 1996) and Fashion-MNIST (Xiao et al., 2017). An overview of the properties of these datasets can be found in Table 1. We chose this diverse set of datasets covering various domains in order to demonstrate the general applicability of our method. On top of that, these datasets are difficult enough such that there is room for improvement.

The MicroMass dataset consists of mass-spectrometry data obtained from bacterial strains. The objective is to discriminate between bacterial species based on spectra, which has been shown to be hard for several species. Note that only pure spectra data was considered. CNAE-9 is a text classification task, in which 1080 documents represented by their word frequencies are classified into 9 categories. This is a sparse dataset: 99.22% of the data consists of zeros. Epilepsy is a time-series dataset from EEG recordings of 500 individuals, in which the objective is to detect epileptic seizures. HAR consists of various smartphone sensor statistics, from which the activity of the person carrying the phone must be deduced. Madelon is an artificial dataset that has 5 informative features and 15 linear combinations of those features. The other 480 features are probes that provide no information about the class label. ISOLET is a speech dataset, from which it must be recognized which letter of the alphabet was spoken by a subject. COIL-100 and Fashion-MNIST are both small grayscale image datasets.

In order to perform experiments on the previously described datasets, they must be split into a training set and a test set. For ISOLET, HAR, Madelon and Fashion-MNIST, such a split was already provided. For Epilepsy, CNAE-9, MicroMass and COIL-100 on the other hand, a custom split had to be made. We opted to randomly sample 20% of the data as test data, the remainder being the training data.

The aim of our experiments is to study the effect of our proposed training algorithms on the accuracy of sparse MLPs, compared to the original approach, i.e. SET-MLP (sparse MLP trained with SET). The sparse MLPs trained with our algorithms are dubbed further: CoDASET-MLP, CoPASET-MLP, CoRSET-MLP, CoDACoRSET-MLP and CoPACoRSET-MLP. Since we are not trying to optimize hyperparameters, we mostly used the same configuration of parameters as SET, which is also our baseline.

For all experiments, the multilayer perceptron that was used consisted of an input layer, three hidden layers of 1000 neurons each and an output layer. The only exception to this are the experiments involving Fashion-MNIST, in which hidden layers of 200 neurons were used instead. In addition, we used activation function SReLU (Jin et al., 2016), sparsity level $\varepsilon = 20$, rewire rate $\zeta = 0.3$ and a dropout

rate of 0.3, all of which were also used in SET. Finally, the learning rate η was chosen by empirically experimenting with different values. MicroMass and Madelon were found to give best results for $\eta = 0.1$, the other datasets use $\eta = 0.01$. Please note that these hyperparameter values yield the same amount of connections for all models studied, i.e. SET-MLP, CoDASET-MLP, CoPASET-MLP, CoRSET-MLP, CoDACoRSET-MLP and CoPACoRSET-MLP, on the same dataset.

4.2. Implementation Details

Our method is implemented on top of Keras, using a weight mask that sets selected weights to zero in order to create a sparse network. The weight rewiring step itself is implemented in pure Python. Proof-of-concept code is available at <https://github.com/joostPieterse/CosineSET>.

4.3. Results

The resulting accuracy plots of our experiments are shown in Fig. 1. Note that the experiments on MicroMass were run multiple times for each method, the plot shown in Fig. 1 is the experiment resulting in median maximum accuracy for that method. The reason for running MicroMass in particular multiple times is twofold. First, when performing experiments on MicroMass, we observed a relatively large variance in accuracy over time. Second, the difference in maximum accuracy between the results of each of the different methods was quite small. Consequently, multiple runs were needed in order to obtain statistically significant results from this dataset.

Table 2 lists the maximum accuracy for each method/dataset combination. Note that the maximum accuracies for MicroMass are averaged over all runs on that dataset/method. Experiments on other datasets are reported after one run, as we did not observe a significant difference at multiple runs. For context, we will first provide an overview of the results of previous research on neural networks on these datasets, before analyzing our own methods. We emphasize, however, that the purpose of these experiments is not necessarily to improve accuracy for a specific dataset by e.g. tuning hyperparameters, but instead to identify the effect on accuracy of integrating our approach into the SET procedure.

To the best of our knowledge, the best neural network approach for ISOLET using no further signal processing has an accuracy of 96.02% (Kochetov and Putin, 2016), by employing an MLP of which the hyperparameters were optimized by a software library. We improve upon this result with an accuracy of 96.54% for our best method, while using only 0.14M parameters compared to their 1.1M parameter MLP. Among the neural network-based approaches to HAR, an accuracy of 95.75% was obtained using a CNN (Ronao and Cho, 2015). Our best result improves this by about 2%, even

Table 1. Dataset statistics

Dataset	Domain	Data type	# classes	# features	# train samples	# test samples
ISOLET	Speech	Continuous	26	617	6238	1559
HAR	Phone sensor	Continuous	6	561	7352	2947
Madelon	Artificial	Discrete	2	500	2000	600
Epilepsy	EEG	Discrete	2	178	9244	2256
CNAE-9	Text	Discrete	9	856	858	222
MicroMass	Mass-spectrometry	Discrete	20	1300	454	117
Fashion-MNIST	Image	Discrete	10	784	60000	10000
COIL-100	Image	Discrete	100	1024	5764	1436

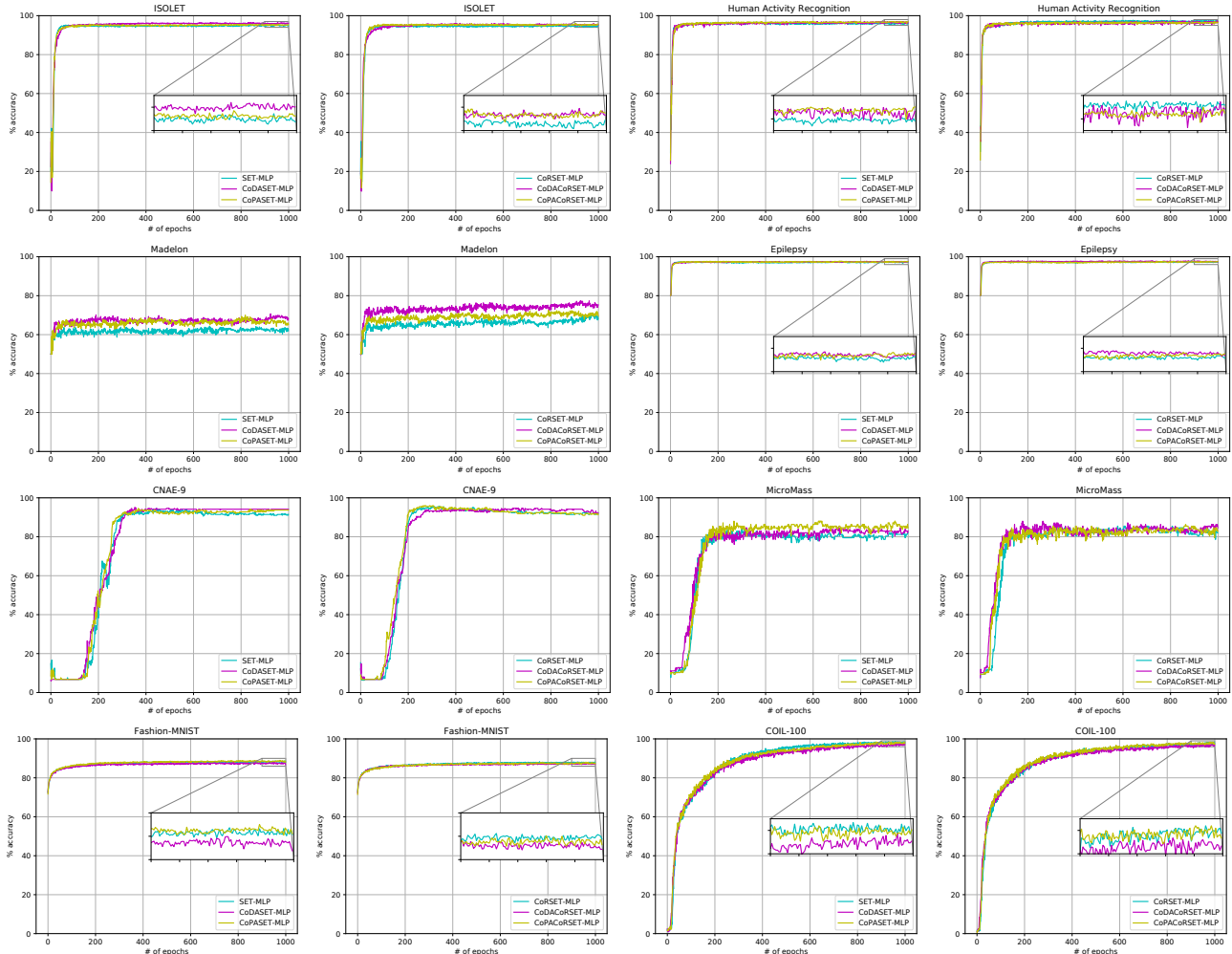


Figure 1. Evaluation of the proposed cosine similarity-based sparse MLP models against the baseline SET-MLP on 8 different datasets.

though the incorporated MLP replacing the softmax layer of their CNN already has twice as many parameters as our MLP. Madelon is not commonly tackled by neural networks. The difficulty of using an MLP for this dataset for example becomes clear in (Santos and Ramos, 2010). They obtained an accuracy of 56.0% for this two-class classification problem when using all features in an MLP, whereas our best method has an accuracy of 77.50%. In (Orhan et al., 2011), the accuracy of recognizing epileptic seizures is improved from 93.2% using an MLP to 99.6% using both an MLP

and K-means. Their result is better than our best accuracy of 97.96%, although this may be because a combination of two techniques is used. For CNAE-9, an accuracy of 97.2% was obtained with an MLP (Ko et al., 2017), which is more than our best result of 95.95% while using a similar number of parameters. An accuracy of 91.06% can be obtained on Fashion-MNIST (Gouk et al., 2018), approximately 2% more than our best result. However, they use a much larger MLP that has 7.0M parameters, while our MLP for Fashion-MNIST only has 0.038M parameters. In (Vicente et al.,

2003), an accuracy of 96% was obtained for COIL-100 by employing both Principle Component Analysis and an MLP, which we improve by over 2%. So, to summarize, the state-of-the-art is improved upon by at least one of the methods listed in Table 2 for 4 out of the 7 previously mentioned datasets, despite the fact that hyperparameters were not optimized for these specific datasets. For the unmentioned dataset, MicroMass, we are not aware of any MLP results reported in the literature. However, in (Vervier et al., 2015) an accuracy of 89.4% was obtained on MicroMass using an SVM-based strategy.

4.3.1. CoDASET-MLP

The accuracy plots of CoDASET-MLP show that its training speed is similar to SET-MLP’s. A notable exception is ISOLET, in which CoDASET-MLP continues to improve accuracy for much longer than SET-MLP. So besides ISOLET, CoDASET-MLP needs a similar number of training epochs to reach maximum accuracy. CoDASET-MLP improves upon the SET-MLP baseline in terms of maximum accuracy in nearly all non-image datasets. On image datasets, however, its performance is relatively poor. An important characteristic of image datasets is that the location of features can shift, as objects appear at different points in the image, which also inspired CNNs. We hypothesize that using cosine similarity makes an MLP less robust to feature shift. However, MLPs are not frequently used for image classification in practice except for benchmarking, as more suitable alternatives such as CNNs exist.

4.3.2. CoPASET-MLP

For almost all datasets, CoPASET-MLP outperforms SET-MLP. The only dataset for which SET-MLP performs better is COIL-100, though it actually achieves highest accuracy on the other image dataset, Fashion-MNIST. The results are also in line with our expectation that its performance would be more consistent compared to CoDASET-MLP across different datasets, since the reintroduced randomness leads to better exploration of possible topologies. Because of these performance improvements, we conclude that this method outperforms SET-MLP in a wide range of applications and gives more consistent results than CoDASET-MLP.

4.3.3. CoRSET-MLP

A small improvement can be observed for most of the non-image datasets compared to SET-MLP, even obtaining best results on HAR over all methods. On the other hand, the accuracies of CoRSET-MLP are in general slightly lower compared to the accuracies of CoDASET-MLP and CoPASET-MLP. So CoRSET-MLP’s results are a small improvement over SET’s, but overall CoDASET-MLP and CoPASET-MLP obtained better results.

4.3.4. CoDACoRSET-MLP

In the experiments on CoDACoRSET-MLP, we observed large improvements on all but one of the non-image datasets. Especially noteworthy are the results on Madelon, Epilepsy and MicroMass, for which the highest accuracy over all methods was obtained showing a clear improvement over SET. The large improvement for Madelon in particular stands out. This is an extremely noisy dataset, so the improved performance may indicate that cosine similarity reduces overfitting on this noise. Furthermore, we can see that the difference in results on image- and non-image data is even more pronounced in the results of this method.

4.3.5. CoPACoRSET-MLP

Here a similar observation can be made: replacing SET’s method of removing connections with CoRSET results in lower performance for image datasets, but generally better results for the other datasets. However, CoDACoRSET-MLP has better results than this method on all but one of the non-image datasets and is therefore a more promising method. Thus we conclude that CoPASET-MLP is more suitable for use on image datasets and CoDACoRSET-MLP is more suitable for non-image datasets.

5. Discussion: Understanding Evolutionary Pattern Behavior

In the results of CoDACoRSET-MLP on Madelon, we noted a large improvement in accuracy over SET-MLP and hypothesized that this indicates that our methods reduce overfitting. In this section, we conduct further analysis of this hypothesis by analyzing the models topologies obtained after training. Herein, we focus just on the Madelon dataset, while an extensive analyze on all datasets can be found in Appendix A. In particular, we are interested in the question of whether this difference is caused by the prioritization of features as the network evolves during training. If the right features are prioritized as the network evolves, their degrees would follow the feature importance distribution.

Madelon has an interesting feature importance distribution. It contains 5 informative features and 15 combinations thereof, for a total of 20 (redundant) informative features. All other 480 features are noninformative noise. The resulting distribution of the degrees is shown in Fig. 2. For SET-MLP, we do observe some input neurons which have more than 60 connections, these are slightly more connected compared to the other neurons. However, for CoDACoRSET-MLP we obtained exactly 20 neurons that are clear outliers (i.e. the input neurons with degree larger than 100). We cannot for sure know if these 20 input neurons correspond to the 20 informative features, as this information is not provided in the dataset. Yet, this is certainly suggested by the

Table 2. Cosine similarity accuracies for our newly proposed algorithms and some combinations thereof. Each entry denotes *accuracy* (relative accuracy compared to SET-MLP), the entry with highest accuracy for its dataset is made bold. Sparsity level represents the number of missing (zero-out) connections in the sparse MLPs from the total number of connections in their corresponding fully-connected MLPs. Per dataset, all sparse MLP models have the same amount of parameters (connections).

DATASET	SET-MLP (%)	CoDASET-MLP (%)	CoPASET-MLP (%)	CoRSET-MLP (%)	CoDACoRSET-MLP (%)	CoPACoRSET-MLP (%)	SPARSITY LEVEL (%)	NUMBER OF PARAMETERS
ISOLET	95.45	96.54 (+1.09)	95.70 (+0.25)	95.13 (-0.32)	96.09 (+0.64)	95.96 (+0.51)	94.76	138K
HAR	96.67	97.12 (+0.45)	97.01 (+0.34)	97.69 (+1.02)	97.46 (+0.79)	97.12 (+0.45)	95.43	117K
MADELON	64.33	70.67 (+6.34)	70.00 (+5.67)	70.67 (+6.34)	77.50 (+13.17)	72.50 (+8.17)	95.52	112K
EPILEPSY	97.47	97.74 (+0.27)	97.78 (+0.31)	97.61 (+0.14)	97.96 (+0.49)	97.65 (+0.18)	95.15	105K
CNAE-9	94.59	95.05 (+0.46)	94.59 (0.00)	95.50 (+0.91)	94.59 (0.00)	95.95 (+1.36)	95.59	126K
MICROMASS	85.47	85.47 (0.00)	87.46 (+1.99)	85.47 (0.00)	88.03 (+2.56)	86.97 (+1.50)	95.60	146K
FASHION-MNIST	88.73	87.97 (-0.76)	89.01 (+0.28)	88.32 (-0.41)	87.62 (-1.11)	88.05 (-0.68)	84.22	37K
COIL-100	98.68	97.77 (-0.91)	98.47 (-0.21)	98.40 (-0.28)	97.35 (-1.33)	97.35 (-1.33)	92.24	220K

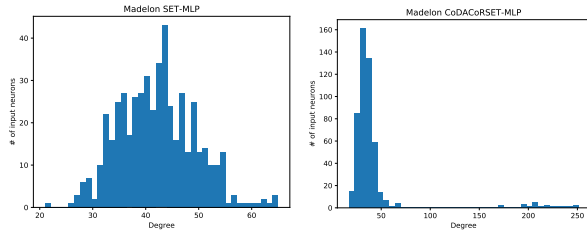


Figure 2. Number of connections distribution per input neuron in SET-MLP and CoDACoRSET-MLP after training on Madelon.

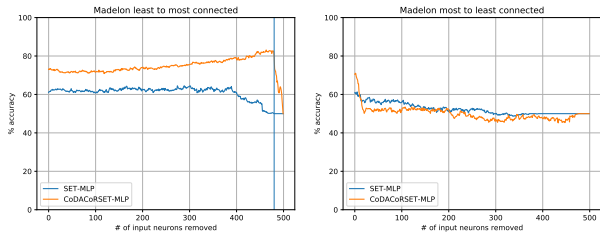


Figure 3. Influence of feature removal on accuracy for the Madelon dataset. The method of selecting features to remove is listed per figure. The vertical line in the left figure marks 480 input neurons removed.

vastly improved results obtained by CoDACoRSET-MLP compared to SET-MLP on this dataset. Furthermore, we show in Fig. 3 the influence on accuracy of removing neurons based on their degree. The left-hand plot shows that we maintain and even improve accuracy while removing 480 of the input neurons that have the least connections in the CoDACoRSET-MLP model. When we continue to remove input neurons after the first 480 neurons, a steep drop in accuracy can be observed, indicating that the 20 outliers in the degree distribution indeed correspond exactly to the informative features in Madelon. The plot for SET-MLP on the other hand shows that its accuracy is also maintained for some time, but clearly more gradually degrades than the plot of CoDACoRSET-MLP. So, we conclude that CoDACoRSET-MLP is a valid approach to understand the impact of input features on the classification task, and may be further use for supervised feature selection and to try understanding the underlying mechanisms of neural networks, while SET-MLP presents also these characteristics but at a

much more diluted level.

6. Conclusion

We introduced a new approach that uses cosine similarity to evolve sparse neural networks. It improves the search process on the optimal topology using domain knowledge. Additionally, based on this approach, five algorithms were proposed to train sparse neural networks, i.e. CoDASET, CoPASET, CoRSET, CoDACoRSET, and CoPACoRSET. All algorithms were tested on 8 different datasets. CoPASET had the most consistent results, outperforming the baseline SET in all but one of the datasets. CoDACoRSET on the other hand performs the best in general on non-image data. In contrast, it obtained the worst results on image data. A possible explanation is the feature shift that can occur in image data. So we can conclude that out of the tested methods, CoDACoRSET is the best method in terms of accuracy for non-image data. It should be noted that for most of the datasets, at least one of the algorithms proposed in this paper outperformed the state-of-the-art on MLPs for that dataset, while frequently having few order of magnitude less connections. Additionally, our experimental results indicate that using cosine similarity for evolution of sparse networks can reduce overfitting. Finally, we show that the evolved connectivity patterns of the input neurons can help understanding the input features impact on classification.

There are several directions for future work. First, further analysis could be conducted on the additional computational complexity introduced by cosine similarity for connection selection. Second, more efficient implementations for all algorithms can be researched, e.g. GPU for cosine similarity (Li et al., 2010), sparse data structures for all neural network models. Third, extensive studies can be performed on the effect of sampling the activation vectors before calculating cosine similarity, which would further reduce its computational complexity and therefore improve scalability. Fourth, trying to understand better the evolved connectivity patterns may lead to more interpretable neural network models.

References

- R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.
- D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium, 2013.
- A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- P. M. Ciarelli and O. Elias. Cnae-9 dataset, 2012. URL <https://archive.ics.uci.edu/ml/datasets/CNAE-9>.
- R. Cole. The isolet spoken letter database. 1990.
- M. Denil, B. Shakibi, L. Dinh, and N. d. Freitas. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems*, pages 2148–2156, 2013.
- S. Dey, Y. Shao, K. M. Chugg, and P. A. Beerel. Accelerating training of deep neural networks via sparse edge processing. In *International Conference on Artificial Neural Networks*, pages 273–280. Springer, 2017.
- H. Gouk, E. Frank, B. Pfahringer, and M. Cree. Regularisation of neural networks by enforcing lipschitz continuity. *arXiv preprint arXiv:1804.04368*, 2018.
- I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the nips 2003 feature selection challenge. In *Advances in Neural Information Processing Systems*, pages 545–552, 2005.
- S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, pages 1135–1143, 2015.
- B. Hassibi, D. G. Stork, and G. J. Wolff. Optimal brain surgeon and general network pruning. In *IEEE International Conference on Neural Networks*, pages 293–299. IEEE, 1993.
- D. O. Hebb. *The organization of behavior: a neuropsychological theory*. Wiley, 1949.
- X. Jin, C. Xu, J. Feng, Y. Wei, J. Xiong, and S. Yan. Deep learning with s-shaped rectified linear activation units. In *Association for the Advancement of Artificial Intelligence*, pages 1737–1743, 2016.
- N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, et al. In-datacenter performance analysis of a tensor processing unit. In *Computer Architecture (ISCA), 2017 ACM/IEEE 44th Annual International Symposium on*, pages 1–12. IEEE, 2017.
- J. H. Ko, D. Kim, T. Na, J. Kung, and S. Mukhopadhyay. Adaptive weight compression for memory-efficient neural networks. In *Proceedings of the Conference on Design, Automation & Test in Europe*, pages 199–204. European Design and Automation Association, 2017.
- K. Kochetov and E. Putin. Specnn: The specifying neural network. In *2016 International Symposium on Innovations in Intelligent Systems and Applications*, pages 1–5, 2016.
- Y. LeCun, J. S. Denker, and S. A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, pages 598–605, 1990.
- Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, May 2015.
- Q. Li, V. Kecman, and R. Salman. A chunking method for euclidean distance matrix calculation on large dataset using multi-gpu. In *9th International Conference on Machine Learning and Applications*, pages 208–213. IEEE, 2010.
- P. Mahé, M. Arsac, S. Chatellier, V. Monnin, N. Perrot, et al. Automatic identification of mixed bacterial species fingerprints in a maldi-tof mass-spectrum. *Bioinformatics*, 30(9):1280–1286, 2014.
- D. C. Mocanu. *Network computations in artificial intelligence - Chapter 5*. PhD thesis, Eindhoven University of Technology, June 2017. URL <https://pure.tue.nl/ws/portalfiles/portal/69949254>.
- D. C. Mocanu, E. Mocanu, P. H. Nguyen, M. Gibescu, and A. Liotta. A topological insight into restricted boltzmann machines. *Machine Learning*, 104(2):243–270, Sep 2016. ISSN 1573-0565. doi: 10.1007/s10994-016-5570-z.
- D. C. Mocanu, E. Mocanu, P. Stone, P. H. Nguyen, M. Gibescu, and A. Liotta. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature Communications*, 2018. doi: 10.1038/s41467-018-04316-3.
- H. Mostafa and X. Wang. Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization, 2019. URL <https://openreview.net/forum?id=S1xBioR5KX>.

- S. Nayar, S. Nene, and H. Murase. Columbia object image library (coil 100). *Department of Computer Science, Columbia University, Technical Report CUCS-006-96*, 1996.
- U. Orhan, M. Hekim, and M. Ozer. Eeg signals classification using the k-means clustering and a multilayer perceptron neural network model. *Expert Systems with Applications*, 38(10):13475–13481, 2011.
- C. A. Ronao and S.-B. Cho. Deep convolutional neural networks for human activity recognition with smartphone sensors. In *International Conference on Neural Information Processing*, pages 46–53. Springer, 2015.
- J. M. Santos and S. Ramos. Using a clustering similarity measure for feature selection in high dimensional data sets. In *10th International Conference on Intelligent Systems Design and Applications*, pages 900–905. IEEE, 2010.
- S. Srinivas, A. Subramanya, and R. V. Babu. Training sparse neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 455–462. IEEE, 2017.
- P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to data mining*, chapter 2, pages 65–76. Pearson Education India, 2006.
- K. Vervier, P. Mahé, J.-B. Veyrieras, and J.-P. Vert. Benchmark of structured machine learning methods for microbial identification from mass-spectrometry data. *arXiv preprint arXiv:1506.07251*, 2015.
- M. A. Vicente, O. Reinoso, C. Perez, C. Fernandez, and J. M. Sabater. Recognition and location of real objects using eigenimages and a neural network classifier. In *Visual Communications and Image Processing 2003*, volume 5150, pages 385–393. International Society for Optics and Photonics, 2003.
- A. Wadhwa and U. Madhow. Bottom-up deep learning using the hebbian principle, 2016.
- H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- H. Zhu and Y. Jin. Multi-objective evolutionary federated learning. *CoRR*, abs/1812.07478, 2018. URL <http://arxiv.org/abs/1812.07478>.

A. Understanding Evolutionary Pattern Behavior

In the results of CoDACoRSET-MLP on Madelon, we noted a large improvement in accuracy over SET-MLP and hypothesized that this indicates that our methods reduce overfitting. In this self-contained appendix, we conduct further analysis of this hypothesis (briefly discussed in Section 5 of the paper) by analyzing the topologies obtained after neural networks training. We investigate whether the differences in accuracy between CoDACoRSET-MLP and SET-MLP can be explained by their network topology. In particular, we are interested in the question of whether this difference is caused by the prioritization of features as the network evolves during training. If the right features are prioritized as the network evolves, their degrees would follow the feature importance distribution. In other words, input neurons that correspond to an important feature would get more connections. Thus, approaches for which this holds can also be used to perform supervised feature selection and to start understanding the behavior of neural networks with adaptive sparse connectivity.

A.1. Method

We analyze the topology of a trained network based on the connectivity of the input neurons. Note that connections added in the final epoch of this trained network were disregarded when calculating the degrees. This is because those connections were not yet optimized by the algorithm. They were only recently chosen, and may have been found to be uninteresting after another training epoch. For SET-MLP, for example, those connections were added randomly and as such are not following the data distribution.

We relate an algorithm’s accuracy to its ability to detect important features by visualizing the degree distribution. If an input neuron has a large degree, then it was considered an important feature during training. By analyzing whether SET-MLP or CoDACoRSET-MLP prioritize the right features, we can determine if these algorithms are suitable for feature selection. If removing scarcely connected input neurons from the network has no significant effect on accuracy, we can conclude that the training algorithm successfully identified that feature as unimportant. Similarly, if removing highly connected input neurons from the network greatly decreases accuracy, then these neurons were important for the network’s performance.

A.2. Results

The same setup as in Section 4 of the paper was used in these experiments as well: an MLP that consisted of an input layer, three hidden layers of 1000 neurons each and an output layer, activation function SReLU (Jin et al., 2016),

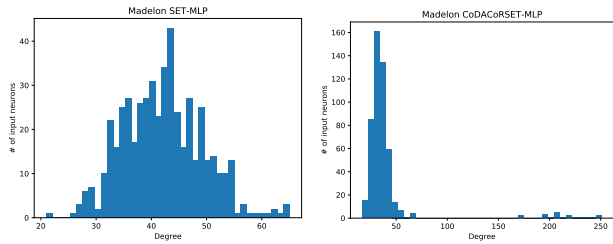


Figure 4. Distribution of the number of connections per input neuron in a SET-MLP and CoDACoRSET-MLP after training on Madelon.

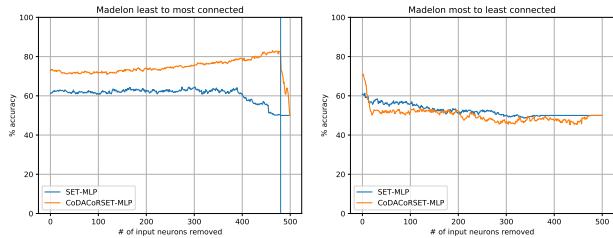


Figure 5. Influence of feature removal on accuracy for the Madelon dataset. The method of selecting features to remove is listed per figure. The vertical line in the left figure marks 480 input neurons removed.

sparsity level $\varepsilon = 20$, rewire rate $\zeta = 0.3$ and a dropout rate of 0.3. We used learning rate $\eta = 0.1$ for MicroMass and Madelon and $\eta = 0.01$ for the other datasets.

A.3. Madelon

First, we analyze the structure of the network after training it on Madelon. In particular, the degree of each input neuron is calculated after training for SET-MLP and CoDACoRSET-MLP. The reason for performing this experiment on Madelon is twofold. First, we would like to understand why there is a particularly large difference in accuracy between SET-MLP and CoDACoRSET-MLP. Understanding how this algorithm improves results for this dataset can help with improving the algorithm, as well as determining in which cases it can be applied. Second, Madelon has an interesting feature importance distribution. It contains 5 informative features and 15 combinations thereof, for a total of 20 (redundant) informative features. All other 480 features are noninformative noise, so their corresponding input neurons should ideally be less connected in the sparse MLP after training. Since this dataset is noisy and provides information on the number of informative features, it is suitable for supervised feature selection experiments.

The resulting distribution of the degrees is shown in Fig. 4. For SET-MLP, we do observe some input neurons which

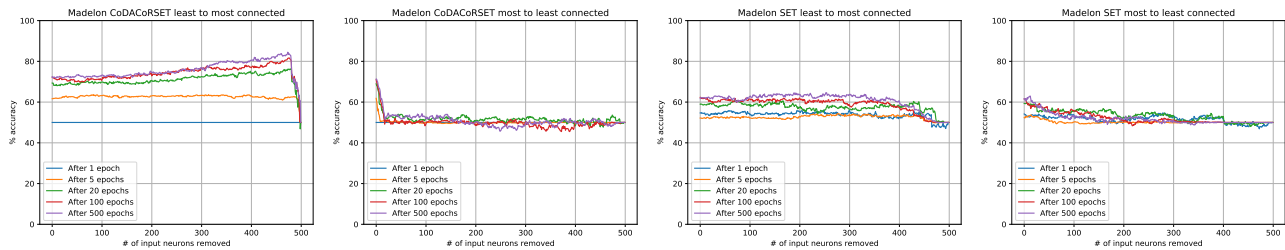


Figure 6. Evolution of feature selection results for Madelon.

have more than 60 connections, these are slightly more connected compared to the other neurons. However, for CoDACoRSET-MLP we obtained exactly 20 neurons that are clear outliers (i.e. the input neurons with degree larger than 100). We cannot for sure know if these 20 input neurons correspond to the 20 informative features, as this information is not provided in the dataset. Yet, this is certainly suggested by the vastly improved results obtained by CoDACoRSET-MLP compared to SET-MLP on this dataset.

Furthermore, we show in Fig. 5 the influence on accuracy of removing neurons based on their degree. The left-hand plot shows that we maintain and even improve accuracy while removing 480 of the input neurons that have the least connections in the CoDACoRSET-MLP network. When we continue to remove input neurons after the first 480 neurons, a steep drop in accuracy can be observed, indicating that the 20 outliers in the degree distribution indeed correspond exactly to the informative features in Madelon. The plot for SET-MLP on the other hand shows that its accuracy is also maintained for some time, but clearly more gradually degrades than the plot of CoDACoRSET-MLP. This indicates that the network of SET-MLP also has highly connected noisy features. From the right-hand plot, which removes the most connected neurons first, we can make the same observation: the accuracy of CoDACoRSET-MLP quickly drops after removing the most connected features, whereas the accuracy of SET-MLP more gradually decreases since it also has highly connected features that are noninformative.

In Fig. 6, feature selection results are shown for several points in the training process. We observe that CoDACoRSET-MLP can already detect Madelon’s informative features after training for only 5 epochs. Thus, it is likely that the additional training epochs do not improve accuracy by improving on the network topology, but mostly by finding better values for the weights. SET-MLP on the other hand does maintain accuracy for longer when given more training epochs. We suspect that SET-MLP needs more training epochs to find the most important features, since it chooses new connections to add to the network randomly and the search space is quite large. On top of that, this plot shows that CoDACoRSET-MLP’s accuracy improves as the

first 480 (noisy) features are removed, indicating that this approach reduces overfitting on the training data.

So, we conclude that CoDACoRSET-MLP is a valid approach to supervised feature selection, while SET-MLP also has some ability to detect relevant features. Furthermore, in the case of Madelon, the difference in accuracy between SET-MLP and CoDACoRSET-MLP can be explained by CoDACoRSET-MLP’s better ability to detect important features.

A.4. Other Datasets

In this section, we perform additional experiments on CoDACoRSET-MLP on non-artificial and less noisy datasets. The reason for executing these experiments is to obtain further insight into CoDACoRSET-MLP’s ability to prioritize important input neurons, such that we can explore whether the difference in accuracy between CoDACoRSET-MLP and SET-MLP can be explained by differences in network topologies. On top of that, it may provide more insight into their feature selection capabilities. A description, statistics and state-of-the-art results of each of these datasets can be found in Section 4. In this section, however, the distribution of the feature importance of these datasets is also discussed. In contrast to Madelon, none of these datasets have features that are purposely informationless. However, they may still be noisy and contain less important features that can be removed with minimum impact on accuracy. For Epilepsy, each sample consists of one second of EEG recording. Since each feature is the same measure at a different point in time, no large differences in feature relevance are expected. ISOLET, MicroMass and HAR all consist of a variety of features, so differences in importance may exist. CNAE-9 and COIL-100 consist of word frequencies and grayscale values respectively, so it is expected that some features are more important than others in these datasets. The feature selection results are shown in Fig. 7 and the corresponding degree distributions are shown in Fig. 8. For most plots we see that accuracy is maintained for some time when removing the least connected features first, indicating that these least connected

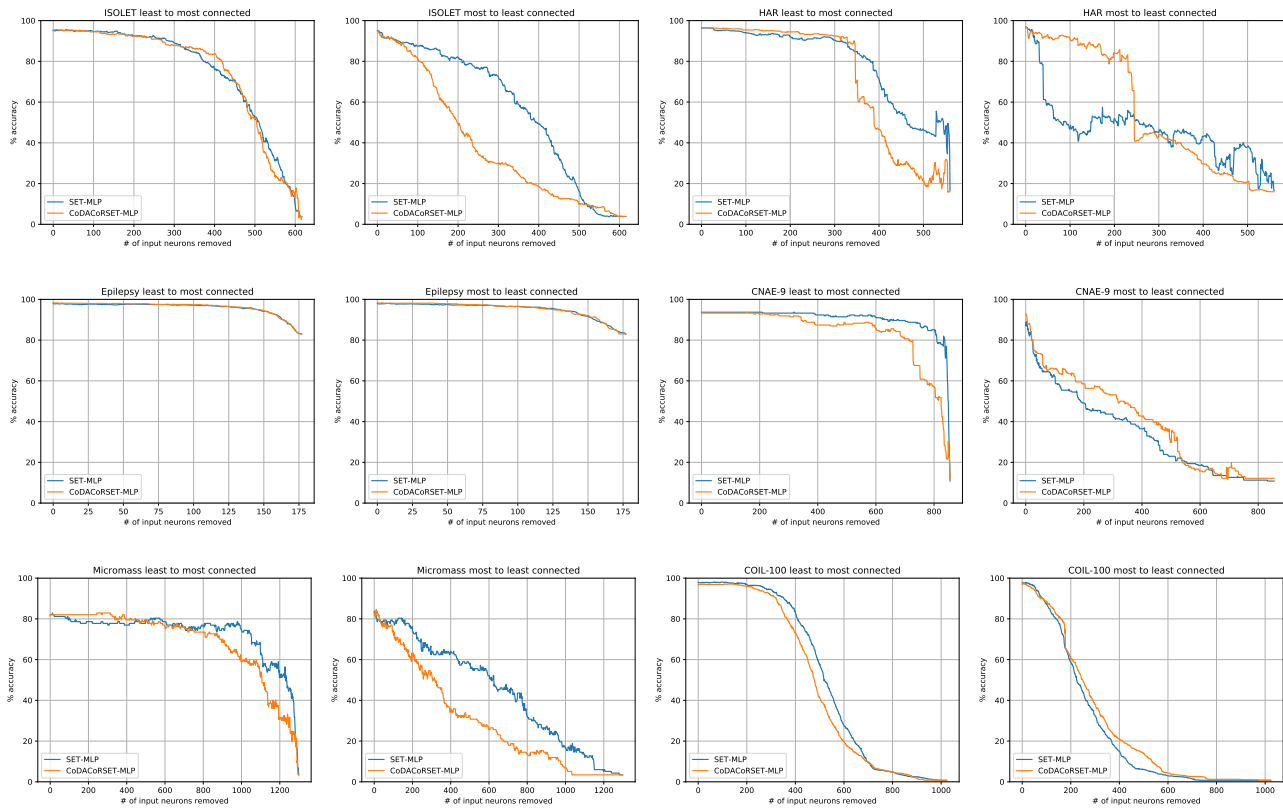


Figure 7. Influence of feature removal on accuracy for the other datasets. The method of selecting features to remove is listed per figure.

features are indeed not essential to prediction performance. We can also see that SET-MLP and CoDACoRSET-MLP show a similar pattern for most datasets, though SET-MLP maintains accuracy for longer for COIL-100. This is in line with the results from Section 4 of the paper, in which SET-MLP performed better on image datasets. On the other hand, CoDACoRSET-MLP maintains accuracy for longer for ISOLET and HAR. This is related to the histograms in Fig. 8, which show that CoDACoRSET-MLP’s degree distribution evolved towards a scale-free distribution for ISOLET and HAR, whereas SET-MLP’s topology did not. A graph has a scale-free topology if the degrees of its nodes follow a power-law distribution (Barabási and Albert, 1999). (Mocanu et al., 2018) suggested that the distribution of the input neurons connections obtained with SET follow the data distribution, but the empirical validation was made just on very few cases. Our results support this claim extensively and show that the distribution of the input neurons connections obtained with CoDACoRSET also follow the data. The results for the other datasets, in which CoDACoRSET-MLP input neurons connectivity did not evolve towards a scale-free distribution is because the degree distribution of the input neurons reflects the features distribution, which for those datasets may not follow a power-law.

To conclude this appendix, we believe that understanding the evolutionary patterns of neural networks with adaptive sparse connectivity, besides the immediate effect of detecting the most impact-full input features on the classification performance, may lead in long term in conceiving more interpretable neural network models.

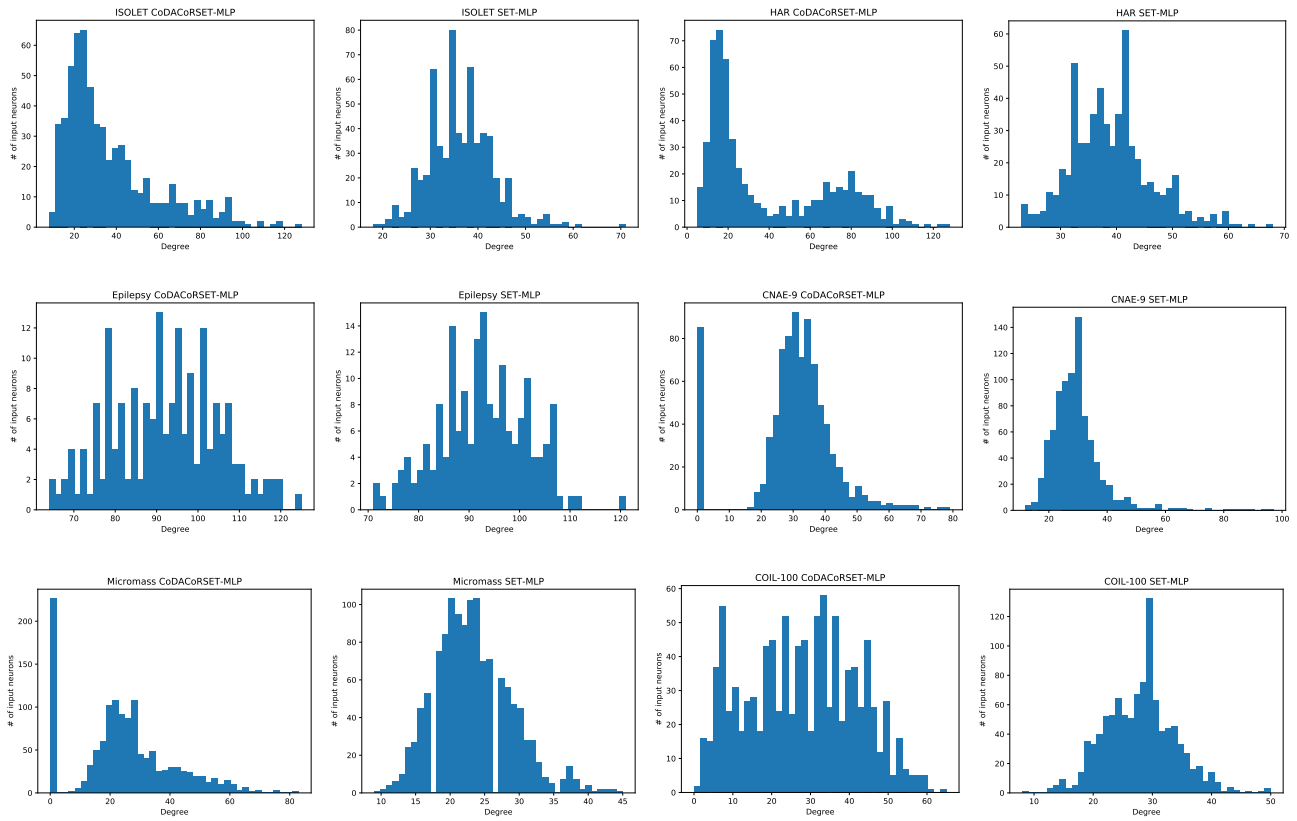


Figure 8. Distribution of the number of connections per input neuron after training the models on the other datasets.