



Lehrstuhl für Data Science

Investigating Sparsity in Recurrent Neural Networks

Masterarbeit von

Harshil Jagadishbhai Darji

1. PRÜFER

Prof. Dr. Michael Granitzer Prof. Dr. Harald Kosch

July 1, 2021

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Research questions	3
1.3	Proposed approach	4
2	Related works	6
3	Technical background	9
3.1	Perceptron	9
3.1.1	Nonlinearity (Nonlinear Activation function)	10
3.1.2	Batch Gradient Descent	13
3.2	Artificial Neural Network	15
3.2.1	Backpropagation	16
3.3	Recurrent Neural Network	20
3.3.1	Backpropagation Through Time	22
3.4	Long Short-Term Memory	25
3.5	Gated Recurrent Unit	27
3.6	Pruning	29
3.6.1	Magnitude-based pruning	29
3.7	Graph Theory	30
3.7.1	Random Graphs	30
3.7.2	Graph properties	31
3.7.3	Small World Networks	33
3.7.4	Scale Free Networks	34
4	Dataset	36
5	Sparse Recurrent Neural Networks	40
5.1	Base model	40

Contents

5.2 Pruned Recurrent Neural Networks	44
5.3 Randomly Structured Recurrent Neural Networks	47
5.3.1 From Random Graph to recurrent network: An example	48
5.3.2 Identifying important graph properties	50
6 Experiment results	52
6.1 Base model performance	52
6.2 Base model performance after pruning	54
6.2.1 Pruning both, input-to-hidden and hidden-to-hidden weights	54
6.2.2 Pruning only input-to-hidden weights	59
6.2.3 Pruning only hidden-to-hidden weights	63
6.3 Performance of Randomly Structured Recurrent Networks	68
6.3.1 Randomly Structured RNN with Tanh nonlinearity	68
6.3.2 Randomly Structured RNN with ReLU nonlinearity	69
6.3.3 Randomly Structured LSTM	70
6.3.4 Randomly Structured GRU	71
6.4 Performance prediction	72
7 Discussion	74
7.1 Base model performance	74
7.2 Pruning recurrent networks	74
7.3 Randomly structured recurrent networks	75
8 Conclusion	77
Appendix A Code	78
Appendix B Tables	79
Appendix C Plots	81
C.1 Randomly structured RNN_Tanh	81
C.2 Randomly structured RNN_ReLU	83
C.3 Randomly structured LSTM	85
C.4 Randomly structured GRU	88
Bibliography	91
Eidesstattliche Erklärung	98

Abstract

In the past few years, neural networks have evolved from simple Feedforward Neural Networks to more complex neural networks, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Where CNNs are a perfect fit for tasks where the sequence is not important such as image recognition, RNNs are useful when order is important such as machine translation. An increasing number of layers in a neural network is one way to improve its performance, but it also increases its complexity making it much more time and power-consuming to train. One way to tackle this problem is to introduce sparsity in the architecture of the neural network. Pruning is one of the many methods to make a neural network architecture sparse by clipping out weights below a certain threshold while keeping the performance near to the original. Another way is to generate arbitrary structures using random graphs and embed them between an input and output layer of an Artificial Neural Network (ANN). Many researchers in past years have focused on pruning mainly CNNs, while hardly any research is done for the same in RNNs. The same also holds in creating sparse architectures for RNNs by generating and embedding arbitrary structures. Therefore, this thesis focuses on investigating the effects of the before-mentioned two techniques on the performance of RNNs. We first describe the pruning of RNNs, its impact on the performance of RNNs, and the number of training epochs required to regain accuracy after the pruning is performed. Next, we continue with the creation and training of Sparse Recurrent Neural Networks (Sparse-RNNs) and identify the relation between the performance and the graph properties of its underlying arbitrary structure. We perform these experiments on RNN with Tanh nonlinearity (RNN-Tanh), RNN with ReLU nonlinearity (RNN-ReLU), GRU, and LSTM. Finally, we analyze and discuss the results achieved from both the experiments.

Acknowledgments

Many people helped and encouraged me during my Master's studies, and I would like to express my gratitude to all of them. However, there are some people whom I would like to thank individually as their vital contribution helped me get this far.

First, I would like to thank Julian Stier, my thesis advisor, for his constant guidance and encouragement to not only focus on my preliminary research questions but also to explore other possibilities to improve and expand my research interests. He was always quick in providing feedback on my work that allowed me to complete my work promptly.

Secondly, I would like to thank my thesis supervisors, Prof. Dr. Michael Granitzer and Prof. Dr. Harald Kosch. Prof. Granitzer's comments on my initial work helped me expand my thesis work. Their comments during the initial presentation of my work helped me think of original and efficient ways to improve my implementations.

During my studies in Passau, I became friends with people who not only made this city a second home, but they were always there whenever I needed them. One of the important things we all learned from this experience is that a friendship goes beyond any borders.

Finally, but most importantly, I would like to thank my parents and my sister for their continuous support and blessings. They always made sure I get the best education, no matter what the situation is. They have always believed in me, and their combined efforts and sacrifices made it possible for me to continue my Master's education in Germany.

List of Figures

1.1	Hand-written digits	1
1.2	Flowchart for the pruning experiment	4
3.1	Single Layer Perceptron	9
3.2	Rectified Linear Unit	11
3.3	Sigmoid activation function	11
3.4	Tanh activation function	12
3.5	Artificial Neural Network	15
3.6	Backpropagation in a neural network	16
3.7	Backpropagating error from output layer	17
3.8	Backpropagating error from h_1	18
3.9	Backpropagating error from h_2	18
3.10	A simple Recurrent Network	20
3.11	A Recurrent Neural Network unfolded in time	20
3.12	Internal structure of a single hidden unit in a standard RNN	21
3.13	Backpropagation Through Time	22
3.14	Internal structure of a single hidden unit in an LSTM	25
3.15	Internal structure of a single hidden unit in a GRU	27
4.1	Flow diagram to generate Reber grammar sequences	36
4.2	Validity and number of strings in train-test dataset	37
4.3	String length distribution of entire dataset	38
4.4	String length distribution of train dataset	39
4.5	String length distribution of test dataset	39
5.1	Training recurrent networks	40
5.2	Generating embeddings of example input sequence	41
5.3	Visualization of a single recurrent layer	42
5.4	Visualization of linear layer's connection to output layer	43

List of Figures

5.5	Simultaneous and individual pruning of i2h and h2h weights	44
5.6	Two densely connected linear layers	45
5.7	Two densely connected linear layers after pruning	46
5.8	Undirected graph containing 5 nodes	48
5.9	Directed graph containing 5 nodes	48
5.10	Neural architecture generated using the layer indexing	49
5.11	Complete model with recurrent architecture generated from a random graph	50
6.1	RNN_Tanh base model performance	52
6.2	RNN_ReLU base model performance	53
6.3	LSTM base model performance	53
6.4	GRU base model performance	54
6.5	RNN_Tanh base model performance after pruning	55
6.6	RNN_Tanh base model performance regain after pruning	55
6.7	RNN_ReLU base model performance after pruning	56
6.8	RNN_ReLU base model performance regain after pruning	56
6.9	LSTM base model performance after pruning	57
6.10	LSTM base model performance regain after pruning	57
6.11	GRU base model performance after pruning	58
6.12	GRU base model performance regain after pruning	58
6.13	RNN_Tanh base model performance after pruning i2h weights	59
6.14	RNN_Tanh base model performance regain after pruning i2h weights . . .	60
6.15	RNN_ReLU base model performance after pruning i2h weights	60
6.16	RNN_ReLU base model performance regain after pruning i2h weights . .	61
6.17	LSTM base model performance after pruning i2h weights	61
6.18	LSTM base model performance regain after pruning i2h weights	62
6.19	GRU base model performance after pruning i2h weights	62
6.20	GRU base model performance regain after pruning i2h weights	63
6.21	RNN_Tanh base model performance after pruning h2h weights	64
6.22	RNN_Tanh base model performance regain after pruning h2h weights . .	64
6.23	RNN_ReLU base model performance after pruning h2h weights	65
6.24	RNN_ReLU base model performance regain after pruning h2h weights . .	65
6.25	LSTM base model performance after pruning h2h weights	66
6.26	LSTM base model performance regain after pruning h2h weights	66
6.27	GRU base model performance after pruning h2h weights	67
6.28	GRU base model performance regain after pruning h2h weights	67

List of Figures

6.29 Performance of the WS and BA based random structure RNN_Tanh	68
6.30 Performance of the WS and BA based random structure RNN_ReLU	69
6.31 Performance of the WS and BA based random structure LSTM	70
6.32 Performance of the WS and BA based random structure GRU	71
C.1 Comparison of basic graph properties and number of layers in WS and BA based RNN_Tanh models	81
C.2 Correlation between test accuracy of RNN_Tanh and its different graph and recurrent network properties - 1	82
C.3 Comparison of basic graph properties and number of layers in WS and BA based RNN_ReLU models	83
C.4 Correlation between test accuracy of RNN_ReLU and its different graph and recurrent network properties - 1	84
C.5 Comparison of basic graph properties and number of layers in WS and BA based LSTM models	85
C.6 Correlation between test accuracy of LSTM and its different graph and recurrent network properties - 1	86
C.7 Correlation between test accuracy of LSTM and its different graph and recurrent network properties - 2	87
C.8 Comparison of basic graph properties and number of layers in WS and BA based GRU models	88
C.9 Correlation between test accuracy of GRU and its different graph and recurrent network properties - 1	89
C.10 Correlation between test accuracy of GRU and its different graph and recurrent network properties - 2	90

List of Tables

4.1	A few examples of true and false Reber sequences	37
4.2	The number of true and false Reber sequences in our train-test set	38
5.1	Hyper-parameters used for training and evaluating the base model	43
5.2	Computing layer indexing of example DAG	49
6.1	Pearson correlation between test accuracy of RNN_Tanh and different graph and recurrent network properties	69
6.2	Pearson correlation between test accuracy of RNN_ReLU and different graph and recurrent network properties	70
6.3	Pearson correlation between test accuracy of LSTM and different graph and recurrent network properties	71
6.4	Pearson correlation between test accuracy of GRU and different graph and recurrent network properties	72
6.5	R-squared values from each regressor algorithm, for each RNN variant	73
B.1	RNN_ReLU - Feature importance scores for the Random Forest regressor under different circumstances	79
B.2	GRU - Feature importance scores for the Random Forest regressor under different circumstances	80

List of Algorithms

1	Standard RNN algorithm	21
2	Compute layer indexing of nodes	47

1 Introduction

A neural network, also known as an Artificial Neural Network (ANN), is an algorithm that mimics the way a human brain operates. It took millions of years of evolution for the human brain to achieve the level of intelligence we observe today. This intelligence helps us quickly interpret and evaluate what we see in our surroundings. For example, consider the following image of hand-written digits from the MNIST dataset [Lec+98]:

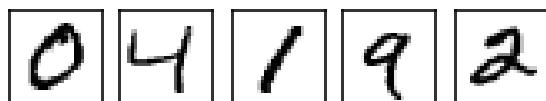


Figure 1.1: **Hand-written digits**: Each hand-written digit is a 28×28 pixel grayscale image. The entire MNIST dataset contains a total of 70000 such images that split into a training set of 60000 images and a test set of 10000 images.

One can quickly recognize these digits as 04192, thanks to the network of billions of neurons available in our brain that makes it easy to identify visual patterns. Past experiences and memories stored in our brain make this process so simple that it happens subconsciously most of the time. This simple process becomes much more difficult when we try to write computer programs to identify similar patterns due to the lack of precise rules and hundreds of exceptions and varieties of a single pattern.

Neural networks tackle this problem by inferring rules from a large set of related training data. For example, to train a neural network to efficiently recognize hand-written digits shown in figure 1.1, a dataset large enough to include different styles of hand-written digits is needed.

Neural networks and their variations such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) are proven to be useful in many different application areas such as image recognition [TKT18], machine translation [BCB14], recommender

systems [PBW17]. Despite this state-of-the-art performance, neural networks are computationally complex and memory-intensive due to their deep structures.

Sparse Neural Networks are a viable option to make neural networks less complex and make them memory efficient. Such sparsity in neural networks can be induced by pruning weights, utilizing skip connections, or generating random architectures based on graphs. Sparsity in a neural network can reduce that network's size by many times with little to no drop in performance [Alf+18; Dey+19; Liu+15]. Therefore, in this thesis, we implement two different ways to achieve sparsity in Recurrent Neural Networks and analyze its performance impact.

1.1 Motivation

As stated before, deep neural networks are likely to have an increased performance but at the cost of higher complexity and increased fast memory requirements. One way to mitigate this issue while maintaining the performance is to introduce sparsity into a network's connections [He+16]. For example, Mao et al. in [Mao+17] report a higher-compression ratio with coarse-grained pruning without loss of accuracy while saving about twice the memory references. Similarly, Sun et al. in [SWT16] reported that the sparse ConvNet model, which only has 12% of the original parameters, still performs similarly to the baseline model.

These are just a few examples where Sparse Neural Networks have proven advantageous in time, energy, and memory savings. Sparsity in traditional neural networks and CNNs is studied widely by many researchers but is not explored much in Recurrent Neural Networks that are difficult to train due to their nonlinear iterative nature. Sparse structures in traditional neural networks have shown a promise of training potential [Alf+18] which, if applied to Recurrent Neural Networks, can also make training RNNs less difficult by reducing their network size while retaining their performance.

1.2 Research questions

The primary research goal of this master thesis is to investigate the effects of sparse structure on the performance of the Recurrent Neural Networks and its variants, namely Long Short-Term Memory and Gated Recurrent Unit. To do so, we have defined the following set of correlated research questions:

1. What is the effect of weights pruning on a recurrent network's accuracy?
2. What percentage of weights pruning is permissible without triggering a significant reduction in the performance?
3. After pruning a certain percent of weights, if we see a significant reduction in the accuracy, how many re-training epochs can regain accuracy?
4. How does a randomly structured recurrent network's performance correlate with the graph properties of its internal structure?
5. Is it possible to predict a randomly structured recurrent network's performance using the graph properties of its base random graph?

We will answer all these questions for four different recurrent networks, namely RNN with Tanh nonlinearity (RNN-Tanh), RNN with ReLU nonlinearity (RNN-ReLU), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), by following the approach explained in section [1.3](#).

1.3 Proposed approach

To investigate the effects of sparse structures in recurrent networks, we propose the following three experiments:

1. Investigating the effects of weights pruning on recurrent networks:

In this experiment, we prune input-to-hidden and hidden-to-hidden weights, both simultaneously and individually.

We start by training a recurrent network on a dataset for a certain number of epochs. Once this training is complete, we follow the process shown in the following flowchart:

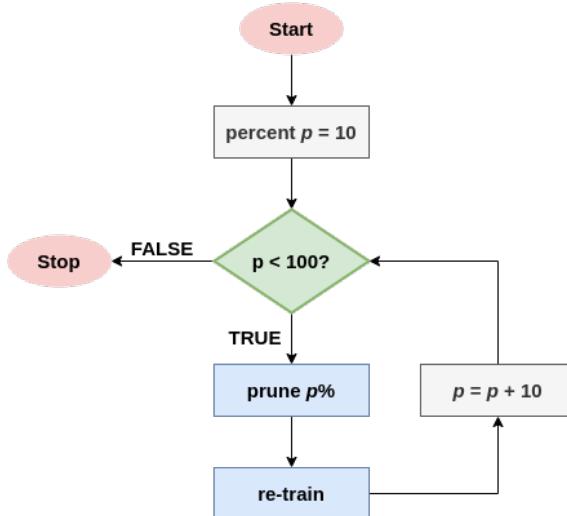


Figure 1.2: The basic flowchart depicting the process of pruning and re-training a trained recurrent network.

To outline this flowchart, once the initial training is complete, retrieve the learned weights of the trained model, prune the lower $p\%$ of weights (where $p \in \mathbb{Z} : p \in [1, 100]$), and re-train this pruned model to check for the number of epochs required to regain accuracy.

This experiment will help answer research questions 1, 2, and 3.

2. Analyzing the performance of recurrent networks with randomly structured recurrent networks:

We start by generating randomly structured neural networks by following the technique described by Stier et al. in [SG19] that generate Sparse Neural Networks.

1 Introduction

After generating Sparse Neural Networks, we introduce recurrent connections to generate Sparse RNNs. These recurrent connections are necessary to make each subsequent run dependent on the previous run.

Once we have Sparse RNNs, we train them on a dataset and compute the correlation between its accuracy and its internal structure's graph properties.

This experiment will help answer the [4th](#) research question.

3. Performance prediction of a randomly structured neural network::

Once we have trained Sparse RNNs, we train three different regressor algorithms with graph properties as features and corresponding accuracy as the target. Since this is a regression problem, we finally report an R-squared value for each RNN variant and each regressor algorithm.

This experiment will help answer the [5th](#) research question.

All these three approaches are described in detail in the later chapters.

2 Related works

A few researchers have worked on pruning to induce sparsity in Recurrent Neural Networks in the past couple of decades. Apart from that, to our knowledge, only one paper explains the process of generating Sparse Neural Networks from randomly generated graphs. This section will summarize these researches to have an idea about what has been done so far.

In September 1994, Lee et al. in [GO94] published one of the first pruning techniques in terms of Recurrent Neural Networks in which authors apply pruning to trained recurrent neural nets. They follow the train, prune, and retrain approach where they train an extensive network on grammar dataset, apply pruning on the trained network, and retrain on the same training set until either satisfactory performance is achieved or the network no longer converge. Since the authors use the incremental training approach, there is no need for the network to use the entire training set, meaning the network can achieve satisfactory achievement using only a part of the training set. After each pruning and retraining cycle, the network would require even less data than the initial network. Based on their experiments' results, the authors claim that “our pruning/retraining algorithm is an effective tool for improving the generalization performance”. According to the authors, this improved performance is due to the reduced size of the network.

In June 2017, Han et al. propose a Recurrent Self-Organising Neural Network (RSONN) in [HZQ17]. To self-organize the RNNs' structure, they use Adaptive Growing and Pruning Algorithm (AGPA). This algorithm works by adding or pruning the hidden neurons based on their competitiveness during the training phase. Authors verify their approach on various benchmark datasets to compare against existing approaches. Based on the experiment results, the authors claim that the proposed RSONN effectively simplifies the network structure and performs better than some exiting methods. Authors compare their performance against various neural network models (such as Fuzzy Neural Network (FNN, [HYJ14]), Exponential Smoothing Recurrent Neural Network (ESRNN,

2 Related works

[YMJ12]), and Self-Organizing Radial Basis Function (SORBF, [HQ12])) and report better performance than others with fewer hidden neurons and less computational complexity. According to the authors, “the proposed RSONN achieved better testing RMSE and running times than the other algorithms”.

In April 2017, a group of researchers from Baidu research published [Nar+17] in which they apply pruning to weights during the initial training of the network. At the end of the training, according to the authors, “the parameters of the network are sparse while accuracy is still close to the original dense neural network”. The authors claim “the network size is reduced by $8\times$ and the time required to train the model remains constant”. As explained in the paper, each recurrent layer has a constant number of hidden units, while we plan to have a different number of hidden units at each recurrent layer. All the experiments are also done on a private dataset, which prevents others from replicating the results, which is necessary to compare this technique of introducing sparsity with other approaches.

In November 2017, another group of researchers from Baidu research published [NUD17] in which they propose a new pruning technique, block pruning. This technique can be used to zero out a block of weights during the training phase of a recurrent network, resulting in a Block-Sparse RNN. In addition to this, authors also use group lasso regularization to check if it can induce block sparsity. Authors report that “block pruning and group lasso regularization with pruning are successful in creating block-sparse RNNs”. Their experiments report a nearly $10\times$ reduction in model size with a loss of 9% to 17% accuracy in vanilla RNN and GRU with block-sparsity of 4×4 blocks. Authors claim their approach is “agnostic to the optimization” and it “does not require any hyper-parameter retuning”.

In September 2019, researchers from the University of Passau published [SG19] in which they aimed to predict the performance of Convolutional Neural Networks (CNNs) using its structural properties. Authors build Sparse Neural Networks (ANNs) by embedding Directed Acyclic Graphs (DAG) obtained through Random Graph Generators into Artificial Neural Networks. Using this approach, they create a dataset of 10000 such graphs, split them into the train-test set, and train it on the MNIST ([Lec+98]) dataset. We will follow a similar approach to obtain arbitrarily structured Sparse Recurrent Neural Networks (Sparse-RNNs). Based on its performance, we will examine the importance of specific structural properties of the internal structure based on its impact on the per-

2 Related works

formance of Sparse-RNNs. This approach can also help in Neural Architectural Search (NAS) for RNNs using performance prediction as a tool.

In November 2019, Zhang et al. proposed a new one-shot pruning approach for Recurrent Neural Networks in [ZS19] obtained from the recurrent Jacobian spectrum. According to the authors, this technique works by “forcing the network to preserve weights that propagate information through its temporal depths”. Authors verified their approach with a GRU network on sequential MNIST ([Lec+98]), Wikitext ([Mer+16]), and Billion words ([Che+13]). Authors claim that “At 95% sparsity, our network achieves better results than fully dense networks, randomly pruned networks, SNIP ([LAT18]) pruned networks, and Foresight ([WZG20]) pruned networks”.

In December 2019, Wen et al. in [Wen+20] explain a different way of applying pruning to speedup Recurrent Neural Networks while avoiding the Lasso-based pruning methods. They introduced two sets of binary random variables to generate sparse masks for the weight matrix that, according to the authors, “act as gates to the neurons”. As explained in the paper, “the presence of the matrix entry w_{ij} depends on both the presence of the i -th input unit and the j -th output unit, while the value of w_{ij} indicates the strength of the connection if $w_{ij} \neq 0$ ”. The optimization of these variables is then performed by minimizing the L_0 norm of the weight matrix. This approach to language modeling and machine reading comprehension works comparatively to the state-of-the-art pruning approaches. Authors claim “nearly $20\times$ practical speedup during inference was achieved without losing performance for the language model on the Penn Treebank dataset”.

Now, after briefly acknowledging the research work done in terms of sparsity in RNNs, in the next chapter, we briefly explain the theoretical background of different approaches and algorithms that we will later use as part of our experiments.

3 Technical background

Before describing pruning and random structures, we must know how RNNs and their different variants, namely LSTM and GRU, operate. Therefore, in this section, we provide a detailed explanation of RNNs, LSTM, and GRU.

To properly understand how RNNs operate, we must understand how a perceptron, a building block of neural networks, works.

3.1 Perceptron

A perceptron, also known as a Single Layer Perceptron (SLP), is a single layer neural network developed by F. Rosenblatt in [Ros58], inspired by [MP43]. It is a linear classifier that accepts multiple binary inputs and returns a single binary output.

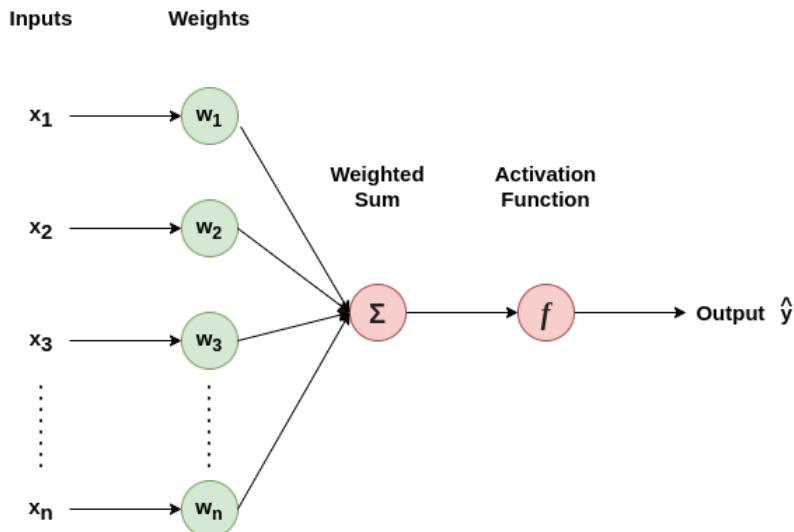


Figure 3.1: **Single Layer Perceptron** with binary inputs $x_1, x_2, x_3, \dots, x_n$ and its corresponding weights $w_1, w_2, w_3, \dots, w_n$.

3 Technical background

As shown in the above figure, a perceptron consists of four main parts, inputs, weights, weighted sum, and an activation function. A perceptron works by following these simple steps:

1. Multiply each input x with its corresponding weight w .
2. Get a value for the weighted sum by adding all the multiplied values together.

$$\text{weighted sum} = \sum_{i=1}^n w_i \cdot x_i \quad (3.1)$$

3. Apply this weighted sum to a activation function to generate the output \hat{y} .

An activation function is a significant part of a perceptron. It transforms the input of the node into the output for that node. It ensures the output value is mapped between $(0, 1)$ or $(-1, 1)$. Rectified Linear Unit (ReLU), Hyperbolic Tangent (Tanh) are two of the popular nonlinear activation functions explained later in the following section.

3.1.1 Nonlinearity (Nonlinear Activation function)

A nonlinearity, as the name suggests, is used when it is not possible to produce an output for any unit using a linear function. Concerning neural networks, three of the most widely used nonlinearities are, ReLU, Sigmoid, Tanh [SLC17].

ReLU

A *Rectified Linear Unit* is a nonlinear activation function mathematically defined as:

$$y = \max(0, x) \quad (3.2)$$

As per the above equation, for a given input x , if the value of x is less than 0, it returns 0, otherwise x .

The following figure shows the line plot of the ReLU activation function.

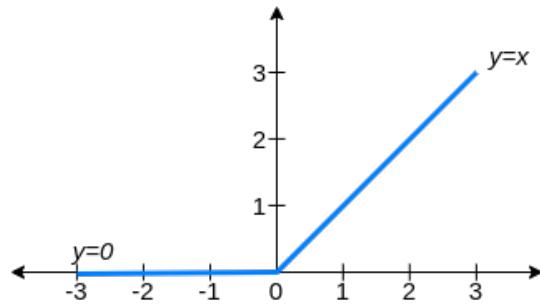


Figure 3.2: Line plot of ReLU activation function

For any given positive input, the derivative of ReLU simply returns 1. This eliminates the need to perform computationally expensive exponentials and divisions, as required by other activation functions such as the Sigmoid activation function.

Sigmoid

The *Sigmoid* activation function, unlike ReLU, is mainly used in *Feedforward Neural Networks*. This activation function returns a value between 0 and 1. The following figure shows the line plot of the Sigmoid activation function.

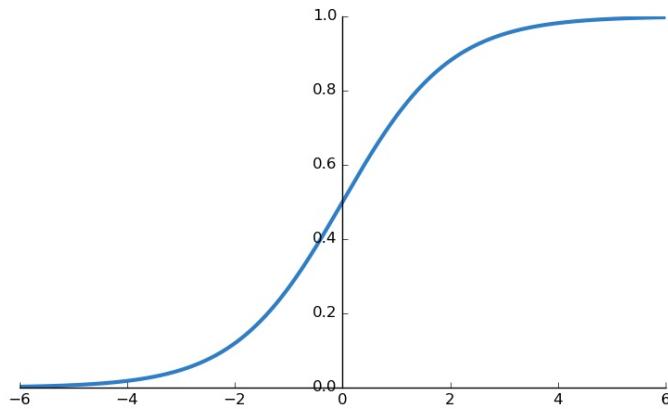


Figure 3.3: Line plot of Sigmoid activation function [Res17]

For a given input x , the Sigmoid activation is mathematically written as:

$$S(x) = \frac{1}{1 + e^{-x}} \quad (3.3)$$

3 Technical background

As mentioned by Nwankpa et al. in [Nwa+18], this activation function has many drawbacks, including gradient saturation and slow convergence. Some of these shortcomings are possible to avoid using other forms of activation functions such as *hyperbolic tangent*.

Tanh

Tanh, short for *hyperbolic tangent* is an activation function with a value range between -1 to 1 , making it a zero-centered activation function. The following figure shows the line plot of the Tanh activation function.

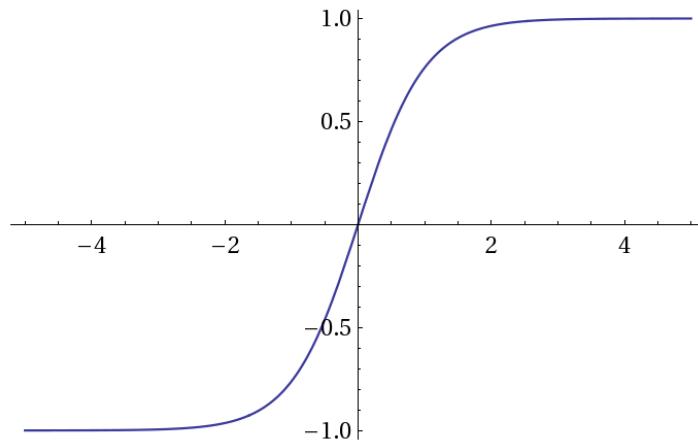


Figure 3.4: Line plot of Tanh activation function [Men18]

For a given input x , the Tanh activation is mathematically written as:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (3.4)$$

Although compared to the Sigmoid activation function, Tanh has better performance during training [KO11; Nea92], it suffers from the vanishing gradient problem.

The output from a nonlinear activation function is not a perfect match to the actual target. For this reason, it is necessary to optimize weight values such that the difference between the actual target and the final output is the smallest, which is done by a process called gradient descent.

3.1.2 Batch Gradient Descent

A perceptron has fixed input and output, meaning we can only modify and improve weights to minimize errors. An error function E returns the deviation of predicted outcome \hat{y} from the actual one y as sum of squared errors:

$$E(w) = \frac{1}{2} \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2 \quad (3.5)$$

Weights are then updated using this error function as:

$$w := w - \eta \nabla E(w) \quad (3.6)$$

where η is the learning rate and $\nabla E(w)$ is partial derivative of the cost function, computed for each weight in the weight vector as:

$$\nabla E(w) = \frac{\partial E(w)}{\partial w_j} \quad (3.7)$$

By substituting the value of $E(w)$ from equation 3.5 in above equation, we can derive $\nabla E(w)$, as shown by [Ras15], as:

$$\begin{aligned} \nabla E(w) &= \frac{\partial}{\partial w_j} \frac{1}{2} \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2 \\ &= \frac{1}{2} \sum_{i=1}^n \frac{\partial}{\partial w_j} (\hat{y}^{(i)} - y^{(i)})^2 \\ &= \frac{1}{2} \sum_{i=1}^n 2(\hat{y}^{(i)} - y^{(i)}) \frac{\partial}{\partial w_j} (\hat{y}^{(i)} - y^{(i)}) \\ &= \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)}) \frac{\partial}{\partial w_j} (\hat{y}^{(i)} - \sum_j w_j \cdot x_j^{(i)}) \\ &= \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)}) (-x_j^{(i)}) \end{aligned} \quad (3.8)$$

3 Technical background

By substituting the value from equations 3.8 in 3.6, we can re-write the weight update formula as:

$$w := w + \eta \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)}) (x_j^{(i)}) \quad (3.9)$$

This approach for updating weights is known as Batch Gradient Descent because each sample in the training set is considered at each step of weights update. Repetition of these training and weights update steps is necessary to obtain convergence.

An SLP has no hidden layers. By adding one or more hidden layers, we can generate a Multi-Layer Perceptron (MLP), also known as an Artificial Neural Network (ANN) or simply, a Neural Network.

3.2 Artificial Neural Network

An Artificial Neural Network is a network of neurons that tries to capture the essential features of the given inputs to infer rules needed to complete a given task, such as image recognition or machine translation. It achieves this by a series of one or more hidden layers, as shown in the below figure:

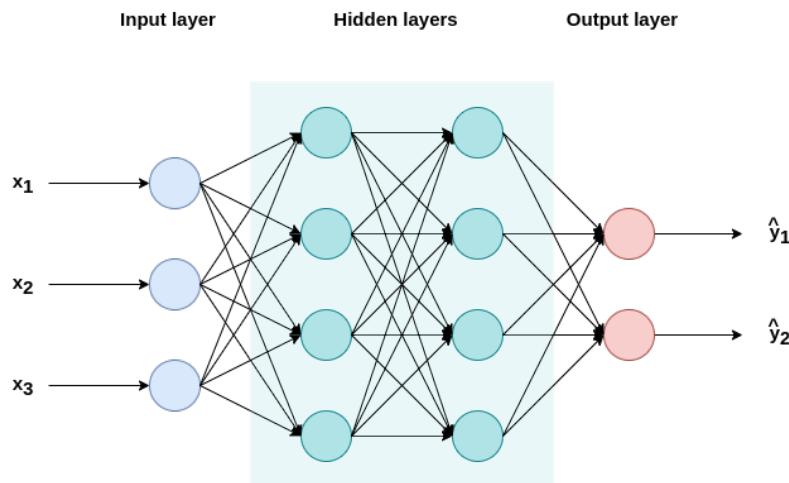


Figure 3.5: **Artificial Neural Network** with two hidden layers, h_1 and h_2 , each consisting of a plethora of neurons.

The neural network shown in the above figure is a Feedforward Neural Network (FFN), where the output of one layer is an input for the next layer. Each neuron of one layer is directly connected to all other neurons of the next layer. Due to this dense connectivity between neurons of each layer, such layers are called dense layers.

As Goodfellow et al. described in [GBC16], the goal of a Feedforward Neural Network is to approximate some function f^* . According to the authors, It does so by defining a mapping $y = f(x; \theta)$ that maps the given input x to a corresponding classification category y . The feedforward network learns the value of θ that returns the best function approximation.

The input layers accept the data from the outside world and transfer it directly to the first hidden layer without performing any computations.

Hidden layers are helpful when the linear separation of the data is not possible. Each neuron in a hidden layer is a perceptron that accepts inputs, computes a weighted sum

(eq. 3.1), applies the weighted sum to an activation function, and passes it towards the next layer.

The output layer of a neural network returns the final results based on the input it receives from its previous layer. The number of neurons in an output layer must match the expected outputs of its respective classification problem. For example, if a neural network's task is to recognize the hand-written digits shown in figure 1.1, then the output layer of this neural network must have ten neurons where each neuron corresponds to a number from 0 to 9.

As with perceptron, neural networks also aim at minimizing the error. Due to the availability of hidden layers, the error is propagated backward using a chain rule, as explained in the following section.

3.2.1 Backpropagation

Although the idea of backpropagation was initially pitched in the 1970s, it became famous by the work of Rumelhart et al. in [RHW86] published in 1986. That paper showed the importance of this algorithm by demonstrating that it works better than other earlier approaches to learning. Given an error function for a neural network, the backpropagation algorithm computes the gradient of this error function. This computation happens backward, from the final layer to the first layer.

To properly understand how chain rule applies in error backpropagation, consider a simple neural network shown in below figure:

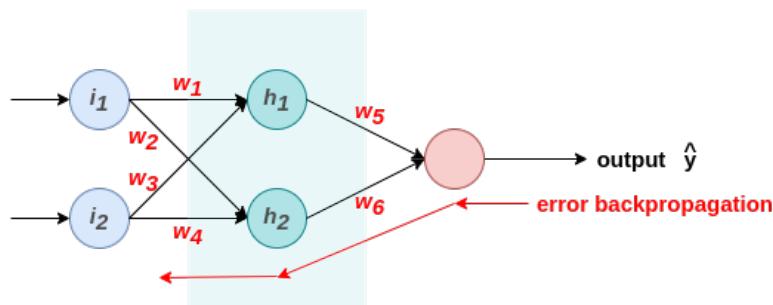


Figure 3.6: **Backpropagation** in a neural network with one hidden layer consisting of two hidden neurons h_1 and h_2 , two input neurons i_1 and i_2 , and one output.

3 Technical background

After getting output from the forward pass, we calculate the error using equation 3.5. By using the partial derivative of this error, the following equation updates all the weights of the neural network:

$$w_j := w_j - \eta \frac{\partial E}{\partial w_j} \quad (3.10)$$

As stated before, we start error backpropagation from the final layer, i.e., the output layer of our neural network shown in figure 3.6. Consider the following figure that is visualizing the output layer of the neural network.

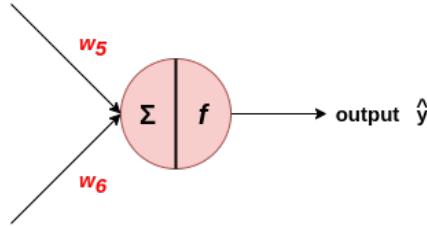


Figure 3.7: Backpropagating error from output layer to modify weights w_5 and w_6 .

To update w_5 and w_6 , we must first compute $\frac{\partial E}{\partial w_5}$ and $\frac{\partial E}{\partial w_6}$. However, error E is the difference between target y and predicted output \hat{y} . Furthermore, the predicted output \hat{y} is the result of applying the weighted sum to an activation function. Here, the weighted sum is calculated as:

$$\text{weighted sum } (z) = \hat{y}_{h1}w_5 + \hat{y}_{h2}w_6 \quad (3.11)$$

where \hat{y}_{h1} is output of the hidden neuron h_1 and \hat{y}_{h2} is output of the hidden neuron h_2 .

Based on this, we can derive the following formula that computes the partial derivative of E with respect to w_5 and w_6 as:

$$\frac{\partial E}{\partial w_5} = \frac{\partial E}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial w_5} \quad (3.12)$$

$$\frac{\partial E}{\partial w_6} = \frac{\partial E}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial w_6} \quad (3.13)$$

3 Technical background

We follow the similar process to compute the partial derivative of error E with respect to w_1 , w_2 , w_3 , and w_4 .

Consider the following figure that is visualizing the hidden neuron h_1 of the neural network:

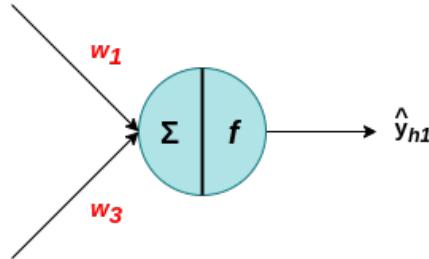


Figure 3.8: Backpropagating error from h_1 to modify weights w_1 and w_3 .

To update w_1 and w_3 , we first compute $\frac{\partial E}{\partial w_1}$ and $\frac{\partial E}{\partial w_3}$ as:

$$\frac{\partial E}{\partial w_1} = \frac{\partial E}{\partial \hat{y}_{h1}} \frac{\partial \hat{y}_{h1}}{\partial z_{h1}} \frac{\partial z_{h1}}{\partial w_1} \quad (3.14)$$

$$\frac{\partial E}{\partial w_3} = \frac{\partial E}{\partial \hat{y}_{h1}} \frac{\partial \hat{y}_{h1}}{\partial z_{h1}} \frac{\partial z_{h1}}{\partial w_3} \quad (3.15)$$

Here, \hat{y}_{h1} is output of the hidden neuron h_1 and z_{h1} is the weighted sum used to compute \hat{y}_{h1} , calculated as:

$$z_{h1} = i_1 w_1 + i_2 w_3 \quad (3.16)$$

Now, consider the following figure that is visualizing the hidden neuron h_2 of the neural network:

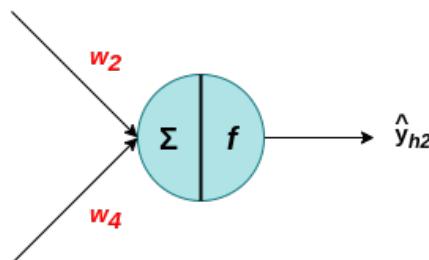


Figure 3.9: Backpropagating error from h_2 to modify weights w_2 and w_4 .

3 Technical background

To update w_2 and w_4 , we first compute $\frac{\partial E}{\partial w_2}$ and $\frac{\partial E}{\partial w_4}$ as:

$$\frac{\partial E}{\partial w_2} = \frac{\partial E}{\partial \hat{y}_{h2}} \frac{\partial \hat{y}_{h2}}{\partial z_{h2}} \frac{\partial z_{h2}}{\partial w_2} \quad (3.17)$$

$$\frac{\partial E}{\partial w_4} = \frac{\partial E}{\partial \hat{y}_{h2}} \frac{\partial \hat{y}_{h2}}{\partial z_{h2}} \frac{\partial z_{h2}}{\partial w_4} \quad (3.18)$$

Here, \hat{y}_{h2} is output of the hidden neuron h_2 and z_{h2} is the weighted sum used to compute \hat{y}_{h2} , calculated as:

$$z_{h2} = i_1 w_2 + i_2 w_4 \quad (3.19)$$

Once we have the partial derivatives of error E with respect to weights, we can use equation 3.10 to update each weight, and then we train the neural network with updated weights. This process of weights update and re-training is repeated until the neural network converges.

Each layer in a traditional neural network is densely connected, making it a complex architecture. Also, in such a neural network, information travels only in the forward direction, i.e., there are no feedback loops available where the input to a function also depends on the output. However, there exist other variations of neural networks that are either computationally less complex (i.e., Convolutional Neural Networks) or support feedback loops (i.e., Recurrent Neural Networks).

A Convolutional Neural Network (CNN) ensures less computational complexity by forcing neurons of one layer to share weights. This sharing of weights reduces the number of learnable parameters, allowing for a better generalization. This process is based on the neocognitron network, published by K. Fukushima in [Fuk80]. Results of [ML15; Lia+16] show that, in the areas of image recognition and classification, CNNs are very useful. However, similar to Feedforward Neural Networks, CNNs also do not have feedback loops. For this reason, Recurrent Neural Networks are better than CNNs while working on tasks where sequence order is essential such as machine translation or music composition.

3.3 Recurrent Neural Network

As stated before, one of the drawbacks of the standard neural network model is the lack of feedback loops. In 1986, M. Jordan published [Jor86], in which he described Recurrent Networks as networks that have a connection from a unit to itself, i.e., a recurrent connection. This recurrent connection act as a feedback loop that makes it possible to use the previous outputs as inputs. This property of Recurrent Neural Networks makes them suitable to work with sequential information where all the inputs depend on each other.

As I. Sutskever describes in [Sut13], a Recurrent Neural Network uses hidden states to incorporate new observations using an intricate nonlinear function. Such a Recurrent Neural Network can effortlessly be understated when it is unfolded in time, as shown in the figure:

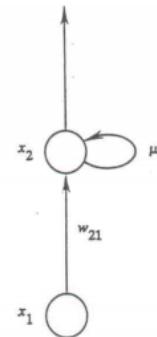


Figure 3.10: A simple Recurrent Network as depicted in [Jor86]

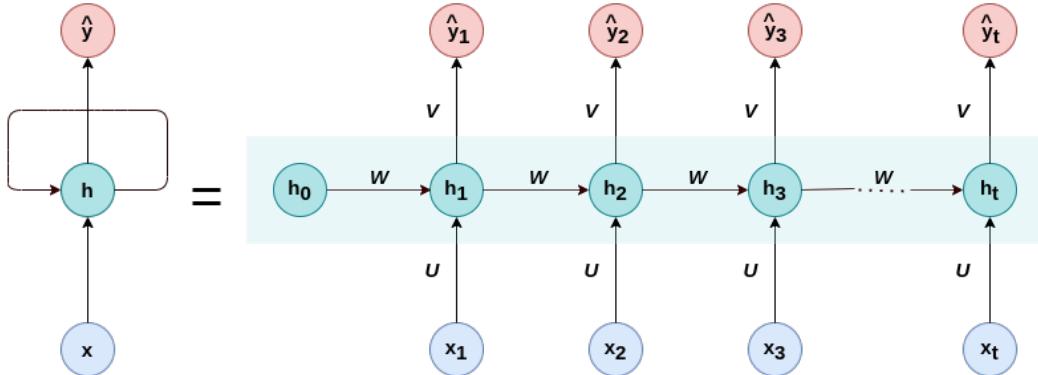


Figure 3.11: A simple **Recurrent Neural Network** unfolded in time with t sequences. Here, U represents *input-to-hidden weights*, W represents *hidden-to-hidden weights*, and V represents *hidden-to-output weights*.

The depicted Recurrent Neural Network accepts inputs $x_1, x_2, x_3, \dots, x_t$, and outputs $\hat{y}_1, \hat{y}_2, \hat{y}_3, \dots, \hat{y}_t$. The hidden states $h_0, h_1, h_2, h_3, \dots, h_t$ are high-dimensional vectors, connected to each other to create recurrence. Deep extensions of a basic RNN can be constructed by stacking multiple recurrent hidden states on top of each other as shown

in [HS13]. Bias vectors b_h and b_o , although are not shown in the above figure, are optional, but are useful to shift the activation function.

The RNN uses the following algorithm to compute h_t and \hat{y}_t :

Algorithm 1: A standard RNN algorithm

```

1 for  $t$  from 1 to  $T$  do
2    $z_t \leftarrow Ux_t + Wh_{t-1} + b_h$            //  $b_h$  is optional
3    $h_t \leftarrow e(z_t)$ 
4    $o_t \leftarrow Vh_t + b_o$                    //  $b_o$  is optional
5    $\hat{y}_t \leftarrow g(o_t)$ 
6 end

```

where $e(\cdot)$ and $g(\cdot)$ are the hidden and output nonlinearities of the RNN. This computation of h_t is visualized in the following figure:

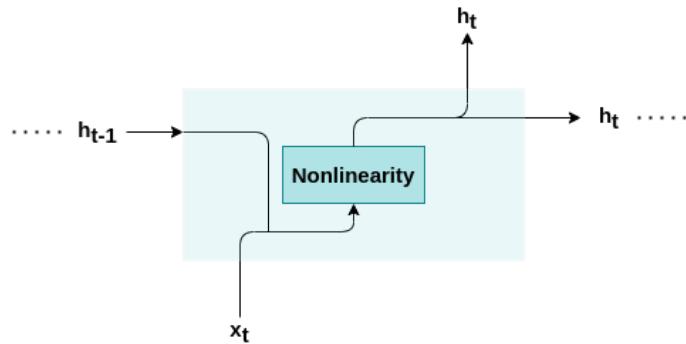


Figure 3.12: The internal structure of a single hidden unit in a standard RNN visualizing the computation of h_t using an input x_t , and the hidden state value of the previous unit h_{t-1} .

The nonlinearity is required to produce a nonlinear decision boundary. Out of all three nonlinearities explained in section 3.1.1, for the scope of this thesis, we only focus on ReLU and Tanh due to their added advantages over the Sigmoid activation function.

Similar to error backpropagation in traditional neural networks (as explained in section 3.2.1), RNNs also backpropagate error to update weights and minimize the absolute error. However, in recurrent networks, since the output of one time-step depends on previous ones, the error also backpropagates from time-step t through the entire network to the first time-step. This is known as Backpropagation Through Time (BPTT [Wer90]).

3.3.1 Backpropagation Through Time

To understand backpropagation through time, consider the following simple RNN with three time-steps visualizing error backpropagation from time-step 3:

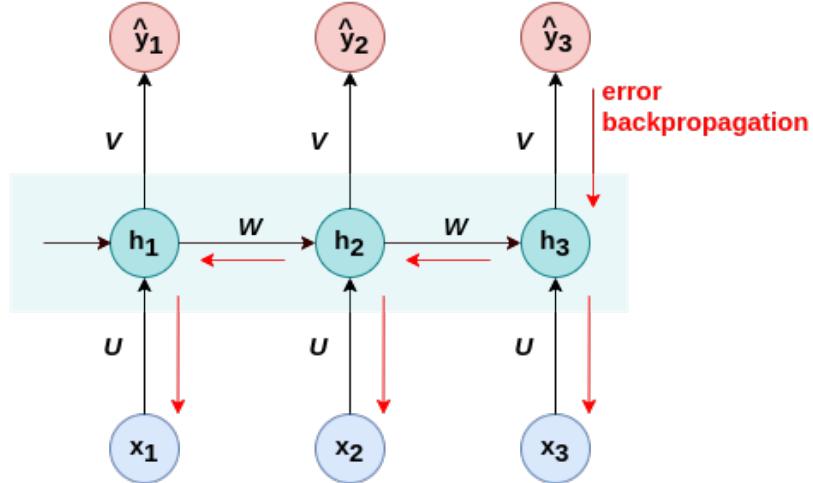


Figure 3.13: A standard RNN visualizing BPTT from time-step 3. Here, U represents *input-to-hidden weights*, W represents *hidden-to-hidden weights*, and V represents *hidden-to-output weights*.

Error at time-step 3 E_3 is the difference between target output y_3 and predicted output \hat{y}_3 . This error E_3 backpropagates from time-step 3 to time-step 1, as shown in the above figure.

To start, using the algorithm 1, we can write following equation for output \hat{y}_3 :

$$\hat{y}_3 = g(o_3) \quad (3.20)$$

where o_3 is product of V and h_3 given as:

$$o_3 = Vh_3 \quad (3.21)$$

where V represents *hidden-to-output weights*, e is output nonlinearity, and h_3 is calculated as:

$$h_3 = e(z_3) \quad (3.22)$$

3 Technical background

where, z_3 is the weighted sum computed as:

$$z_3 = Ux_3 + Wh_2 \quad (3.23)$$

where U is *input-to-hidden* weights, W is *hidden-to-hidden* weights, and e is hidden nonlinearity.

As we can see in the above equations, to compute output \hat{y}_3 , we need a total of three different weight vectors V , W , and U . Since our error E_3 is dependent on output \hat{y}_3 , we must calculate the partial derivative of the error E_3 with respect to all three weight vectors, i.e., we need to compute $\frac{\partial E_3}{\partial V}$, $\frac{\partial E_3}{\partial W}$, and $\frac{\partial E_3}{\partial U}$. To compute these gradients, we follow the same chain rule as we did in backpropagation (section 3.2.1).

Calculating $\frac{\partial E_3}{\partial V}$ is easy as it only depends on \hat{y}_3 , o_3 . Therefore, by following the simple backpropagation, we can calculate $\frac{\partial E_3}{\partial V}$ as:

$$\frac{\partial E_3}{\partial V} = \frac{\partial E_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial o_3} \frac{\partial o_3}{\partial V} \quad (3.24)$$

Calculation of $\frac{\partial E_3}{\partial W}$ is where calculation gets more complicated. To see why it is complicated from $\frac{\partial E_3}{\partial V}$, we first apply the chain rule as:

$$\frac{\partial E_3}{\partial W} = \frac{\partial E_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial h_3} \frac{\partial h_3}{\partial W} \quad (3.25)$$

However, as we can see in equation 3.23, z_3 depends on h_2 , which depends on h_1 . For this reason, we cannot treat h_2 as constant; instead, we need to sum up the contributions of h_2 and h_1 as:

$$\begin{aligned} \frac{\partial h_3}{\partial W} &= \frac{\partial h_3}{\partial W} + \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} \\ &= \frac{\partial h_3}{\partial h_3} \frac{\partial h_3}{\partial W} + \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial h_3}{\partial h_1} \frac{\partial h_1}{\partial W} \\ &= \sum_{t=1}^3 \frac{\partial h_3}{\partial h_t} \frac{\partial h_t}{\partial W} \end{aligned} \quad (3.26)$$

3 Technical background

By substituting value from equation 3.26 in equation 3.25, we get

$$\frac{\partial E_3}{\partial W} = \frac{\partial E_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial h_3} \left[\sum_{t=1}^3 \frac{\partial h_3}{\partial h_t} \frac{\partial h_t}{\partial W} \right] \quad (3.27)$$

We follow the similar process to calculate $\frac{\partial E_3}{\partial U}$ as:

$$\frac{\partial E_3}{\partial U} = \frac{\partial E_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial h_3} \left[\sum_{t=1}^3 \frac{\partial h_3}{\partial h_t} \frac{\partial h_t}{\partial U} \right] \quad (3.28)$$

Similar to the backpropagation algorithm, we use these partial derivatives of error with respect to V , W , and U to update the corresponding weights.

In standard RNNs, each hidden unit only performs the specified nonlinear activation (as depicted in figure 3.12). However, there exist other variations of RNNs, such as Long Short-Term Memory, Gated Recurrent Unit, that do more than just performing a single computation per hidden unit.

3.4 Long Short-Term Memory

Two of the problems with this algorithm are of exploding gradients¹ or vanishing gradient². Two mitigate these problems, in 1997, Hochreiter et al. proposed a new recurrent architecture, termed Long Short-Term Memory (LSTM) in [HS97].

As opposed to the standard RNN's internal structure (figure 3.12), LSTM has a very different and much more complex internal structure made up of three gates (i.e., an *input gate* i_t , an *output gate* o_t , and a *forget gate* f_t) to regulate the flow of information [VMN20]. This is visualized in the following figure:

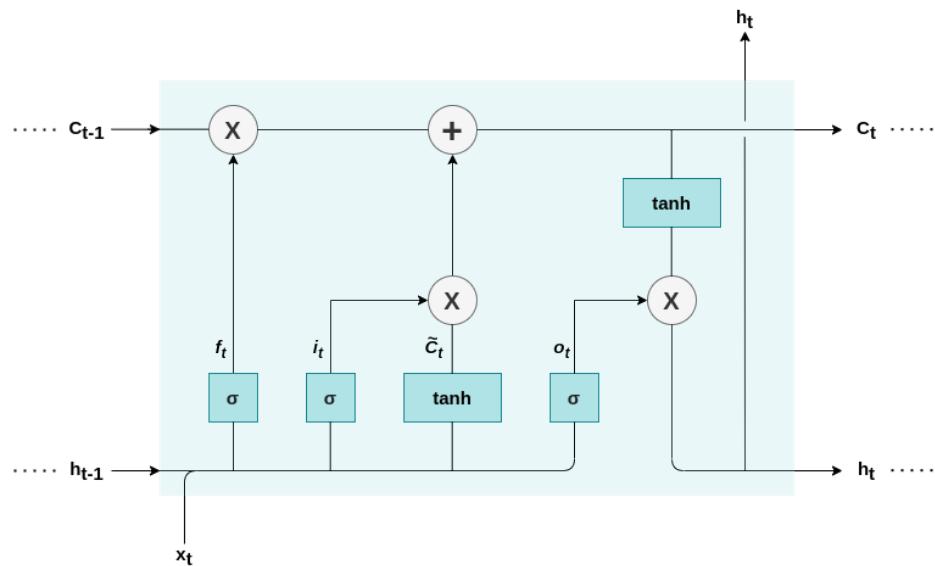


Figure 3.14: The internal structure of a single hidden unit in an LSTM visualizing the computation of h_t and C_t using an input x_t , the hidden state value of the previous unit h_{t-1} , and the cell state value of the previous unit C_{t-1} .

Given an input x_t , previous hidden state h_{t-1} , and previous cell state C_{t-1} , the current hidden state h_t and cell state C_t can mathematically be calculated as follow:

1. The first step is to compute the forget gate that informs the cell state about which past information to keep and which one to forget.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3.29)$$

¹Exploding gradients is a problem when large error gradients accumulate, resulting in substantial updates to model weights making the model unstable.

²The vanishing gradient is a problem that prevents making changes in the weight values due to vanishingly small gradients.

3 Technical background

For each value in the cell state C_{t-1} , it returns a value between 0 and 1.

2. The second step is to determine what new information needs to be stored in the cell state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3.30)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3.31)$$

The addition of these two values then will be used to update the previous cell state.

3. The third step is to update the previous cell state as:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (3.32)$$

4. The final step is to compute the value for the output gate, that then will be used to update the hidden state as:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (3.33)$$

$$h_t = o_t * \tanh(C_t) \quad (3.34)$$

These updated state values, h_t , and C_t are then forwarded to be used in the next unit.

One of the benefits of LSTM over a standard RNN is the capability of learning long-term dependencies. Because of this added benefit, LSTMs have shown exceptional performance in many application areas including time series forecasting [STS18], drug design [Gup+18], and music composition [ES02].

Although LSTMs have a better performance than standard RNNs, their execution is slower due to the availability of a higher number of computations. Therefore, another variation of RNNs, the Gated Recurrent Unit, is faster than an LSTM but has a better performance than a standard RNN.

3.5 Gated Recurrent Unit

The Gated Recurrent Unit (GRU) was introduced by Cho et al. in [Cho+14]. Similar to Long Short-Term Memory, Gated Recurrent Unit also aims to solve the vanishing gradient problem.

The internal structure of a GRU is very different from that of a standard RNN unit but is quite similar to that of an LSTM unit, except with only two gates (i.e., an *update gate* z_t , and a *reset gate* r_t) than three. These two gates control the flow of information that flows into and out of the memory [GG16]. This internal structure is visualized in the following figure:

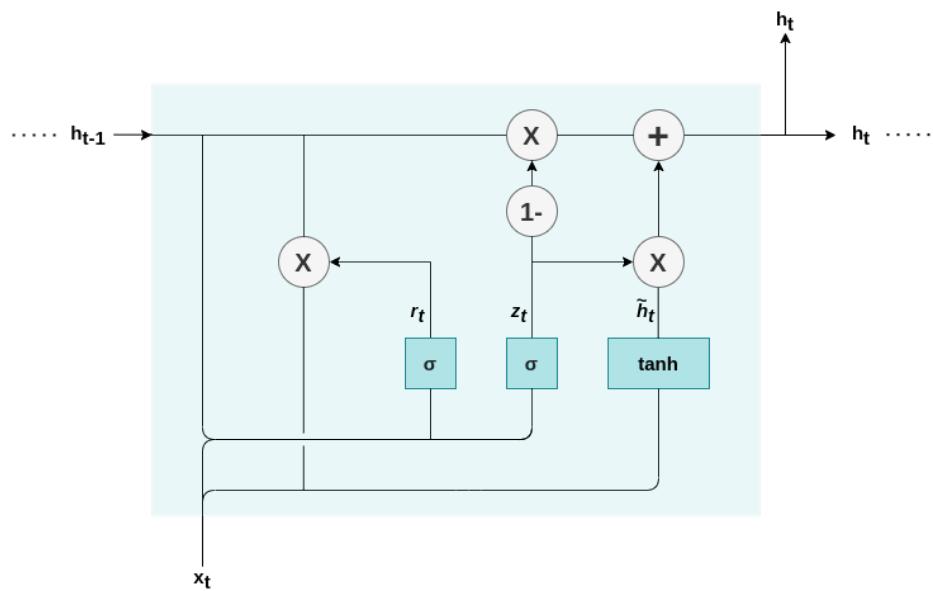


Figure 3.15: The internal structure of a single hidden unit in a GRU visualizing the computation of h_t using an input x_t , and the hidden state value of the previous unit h_{t-1} .

Given an input x_t , and previous hidden state h_{t-1} , the current hidden state h_t can mathematically be calculated as follow:

1. The first step is to compute the update gate value as:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \quad (3.35)$$

This gate helps to determine what information from the past to carry forward.

3 Technical background

2. The second step is to compute the value of the reset gate as:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (3.36)$$

This gate decides how much of the past information to omit.

3. The final step is to update the hidden state using the values of the update gate and the reset gate obtained from equations 3.35 and 3.36, respectively.

$$\tilde{h}_t = \tanh(W_h \cdot [r_t * h_{t-1}, x_t] + b_h) \quad (3.37)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (3.38)$$

This updated hidden state value h_t is then forwarded to be used in the next unit.

Although LSTMs mostly outperform GRUs in many scenarios, GRUs have the upper hand when the dataset is small [YZZ20]. Similar to LSTM, GRUs are also capable of learning long-term dependencies. Because of this, GRUs have shown outstanding performance in different application areas, including speech recognition [Rav+18] and water level prediction [LHL20].

In the scope of this thesis, we only focus on working with RNN with Tanh nonlinearity (RNN-Tanh), RNN with ReLU nonlinearity (RNN-ReLU), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) as these four are the most popularly practiced recurrent networks.

3.6 Pruning

It is well known that the use of more extensive neural networks is favorable to achieve better performance, but at the same time, they are more expensive to use, takes more time to run, and in most cases, require expensive hardware. Therefore, in the past couple of decades, many researchers have worked on various model compression techniques to train large neural networks more efficiently while maintaining their accuracy. Pruning is one such technique that removes the top k ranked weight parameters of a neural network. The weight parameters removed first are usually the ones with the most negligible impact on the model's final output.

In [Bla+20], Blalock et al. studied 81 different papers on pruning to identify important (hyper-)parameters and methods that give the best results with pruning compared to others. In that paper, the authors state that “pruning methods vary primarily in their choices regarding sparsity structure, scoring, scheduling, and fine-tuning”. Here, the sparse structure results from performing unstructured pruning in which individual weight parameters are pruned, and fine-tuning is the process of retraining the pruned neural network to identify the amount of pruning that can be applied without significantly reducing the performance.

There are different pruning techniques such as Magnitude-based pruning, Error-based pruning, Entropy-based pruning, Evolutionary pruning. In the scope of this thesis, we only focus on Magnitude-based pruning as it is proven to give satisfactory results despite being an easy technique.

3.6.1 Magnitude-based pruning

This is the simplest way of pruning that considers the magnitude of weight parameters as pruning criteria, meaning weight parameters that are least important are zeroed out before retraining the model to perform fine-tuning. Li et al. in [Li+18] state the OLMP (**O**ptimization based **L**ayer-wise **M**agnitude-based **P**runing) can reduce the size of AlexNet by 82% without any loss inaccuracy. We also apply layer-wise pruning, where we first calculate threshold based on the percent of pruning to apply and zero-out values below this threshold using a binary mask.

3.7 Graph Theory

Graph theory is the study of graphs, a network of nodes connected with edges. In [Car20], Carson states that the graph theory originated from Leonhard Euler's solution to the famous Königsberg bridge problem. This section briefly explores random graphs, their graph properties, and two random graphs with small-world properties, namely Watts–Strogatz and Barabási–Albert.

An undirected graph G is given as a set of (V, E) , where V is the set of nodes, and E , where $E \subseteq E_{comp} := \{\{v, w\} | v, w \in V, v \neq w\}$ is the set of edges. A complete graph, G_{comp} , is a graph where each node is connected to every other node in the same network, which is different from a connected graph, where it is possible to get from one node to any other node in the same network via a series of edges.

Graphs, in mathematics, are usually used to model real-world structures, but if the structure is very complex and cannot model it in all details, random graphs are used.

3.7.1 Random Graphs

Random graphs, as the name states, have randomly distributed edges according to some probability measure. Based on which adequate probability measure is used, random graphs can be given as:

1. **Uniform random graph:** For the given set of nodes V , edges M , and a probability space Ω where all graphs with nodes V and M edges has the uniform distribution, the probability is given as:

$$P(G) = \binom{M_{comp}}{M}^{-1} \quad (3.39)$$

2. **Binomial random graph:** Given the probability $0 \leq p \leq 1$ that there exists an edge between two given nodes $v, w \in V$ and a probability space Ω where all graphs with nodes V and edges M has the binomial distribution, the probability is given as:

$$P(G) = p^M \cdot (1 - p)^{M_{comp} - M} \quad (3.40)$$

3.7.2 Graph properties

This section explores graph properties that we later use to find a correlation with performance. Along with simple properties such as the number of layers, the number of nodes, and the number of edges, we also use other graph properties, as given below:

Diameter

A graph's diameter is the maximal distance between any pair of nodes in a given graph, excluding any detours or loops. We can use the following equation to find the graph's diameter:

$$\delta = \max_{ij} \{s(i, j)\} \quad (3.41)$$

where $s(i, j)$ is the shortest path between two nodes i and j .

Density

The density of a graph is the ratio of edges present to all possible edges. Given an undirected graph, the density of this graph is calculated as:

$$d = \frac{2m}{n(n - 1)} \quad (3.42)$$

For a directed graph, the density is calculated as:

$$d = \frac{m}{n(n - 1)} \quad (3.43)$$

where n is the number of nodes, and m is the number if edges.

Average Shortest Path Length

The average shortest path length is the mean number of steps on all potential pairs of nodes' shortest paths and is calculated as:

$$a = \sum_{i,j \in V} \frac{d(i, j)}{n(n - 1)} \quad (3.44)$$

3 Technical background

where V is the set of nodes, $d(i, j)$ is the shortest path from node i to j , and n is the number of nodes.

Eccentricity

The eccentricity of a given node V in a connected graph G is the maximum distance from node V to all other nodes in that graph. In contrast, for a disconnected graph, all nodes have an infinite eccentricity.

The maximum a graph can have is the graph diameter, while the minimum eccentricity is the graph radius.

Degree

The degree of a node is the number of nodes adjacent to that node.

Closeness

Closeness, often known as closeness centrality, indicates how close a node is to all other nodes in a given graph and is calculated as:

$$C(u) = \frac{n - 1}{\sum_{i=1}^{n-1} d(i, j)} \quad (3.45)$$

where $d(i, j)$ is the shortest path between nodes i and j , n is the number of nodes in the graph.

Node betweenness

Node betweenness, also known as the betweenness centrality of node u , is the sum of the fraction of all shortest path pairs that pass through u and is calculated as:

$$c_B(u) = \sum_{i,j \in V} \frac{\sigma(i, j|u)}{\sigma(i, j)} \quad (3.46)$$

3 Technical background

Here, V is the set of nodes, $\sigma(i, j)$ is the number of shortest paths between nodes i and j , and $\sigma(i, j|u)$ is the number of shortest paths through node u .

Edge betweenness

In contrast to node betweenness, Edge betweenness is the number of shortest paths that pass through an edge in a given graph. It is calculated as:

$$c_B(e) = \sum_{i,j \in V} \frac{\sigma(i, j|e)}{\sigma(i, j)} \quad (3.47)$$

Here, V is the set of nodes, $\sigma(i, j)$ is the number of shortest paths between nodes i and j , and $\sigma(i, j|e)$ is the number of shortest paths through an edge e .

3.7.3 Small World Networks

In a Small World Network, the mean shortest-path distance between any two given nodes increases sufficiently slowly as a function of the number of nodes in the network [Por12].

A Small World Network have the following three properties:

1. **Higher order of the graph:** The order of a graph is simply the number of nodes in that graph, in other words, the cardinality of its node-set. In a Small World Network, the order of the graph is much higher than its average vertex degree.
2. **Small characteristic path length:** As stated by F. Schreiber in [Sch13], the characteristic path length of a given graph is the average number of edges in the shortest paths between all pairs of nodes. For a given graph to be considered a Small World Network, it must have a small characteristic path length.
3. **High clustering coefficient:** A clustering coefficient is a measure of the degree to which nodes in a graph are clustered together. A given graph must have a higher clustering coefficient to be considered a Small World Network.

Watts–Strogatz is one of the popular Small World Networks, that is explained in the following section.

Watts–Strogatz model

Watts–Strogatz (WS) are regular graphs with degree k . In 1998, Watts and Strogatz [WS98] proposed the Watts-Strogatz model to counteract the incomplete regularity and randomness in social networks.

The construction of a WS model begins with a lattice structure³ that has a high clustering coefficient (C) but also a large characteristic path length (L). We can rewire some edges in this lattice structure to reduce L .

For each node $i \in V$, for each edge $\{i, j\}$ from node i to one of the r right neighbor j , the edge is left unchanged with probability $1 - p$, and with probability p , the edge is replaced by a new edge $\{i, j'\}$ where $j' \in V \setminus (N(v) \cup \{v\})$ is randomly chosen.

For a small probability p , a sharp drop in characteristic path length (L) can be observed with no change in clustering coefficient (C).

3.7.4 Scale Free Networks

In contrast to Small World Networks, where graphs have a small path length due to local clustering, Scale Free Networks have a skewed degree distribution [ANH13]. Probability distribution of node degrees in such graphs is given as:

$$P(d(v) = k) \sim k^{-\gamma} \quad (3.48)$$

where $d(v)$ is the degree of node $v \in V$ and γ is the power law factor.

As seen in the above equation, this probability distribution has power law where the probability of a node having a degree k varies as the power law factor γ varies.

In a Scale Free Network, the system grows with time, and new nodes and edges are added based on the preferential attachment. Such networks are observed in different science and technology areas such as the internet (nodes are HTML pages and edges are hyperlinks), electronic mail (nodes are email addresses and edges are sent emails between two users), and scientific literature (nodes are publications and edges are citations).

³A lattice graph is a graph embedded in a Euclidean space \mathbb{R} that has a regular tiling form [Wei].

Barabási–Albert model

Barabási–Albert model is one of many Scale Free Networks that follows preferential attachment with power-law growth for a given system.

The construction of a Barabási–Albert model begins with $N_0 > 1$ nodes and $m < N_0$ edges where $N_0 = |V_0|$, cardinality of set of nodes at $t = 0$ without any edges.

At time $t = 1$, a new node is added to V_0 with m edges from this new node to m random different nodes.

At time $t > 1$, a new node is added to V_{t-1} with m edges from this new node to m random different nodes, where the preferential attachment gives the probability of choosing node i as:

$$P(i) = c \cdot d(i) \quad (3.49)$$

$$\text{where, } c = \frac{1}{\sum_{j \in V_{t-1}} d(j)}$$

In any given Barabási–Albert model, for any given time t , there are $N_0 + t = N_t$ nodes and $m \cdot t$ edges available.

4 Dataset

For our experiments, we make use of a dataset composed of Reber sequences. A Reber sequence is a grammar string made of finite states, in simple words, formed by using a confined set of characters. In the research paper that proposed the LSTM [HS97], the authors use embedded Reber grammar due to its short time lags. For our purpose, we use a simple version of this embedded Reber grammar, called Reber grammar ([Reb76]).

The following figure shows the flow diagram to generate Reber grammar sequences:

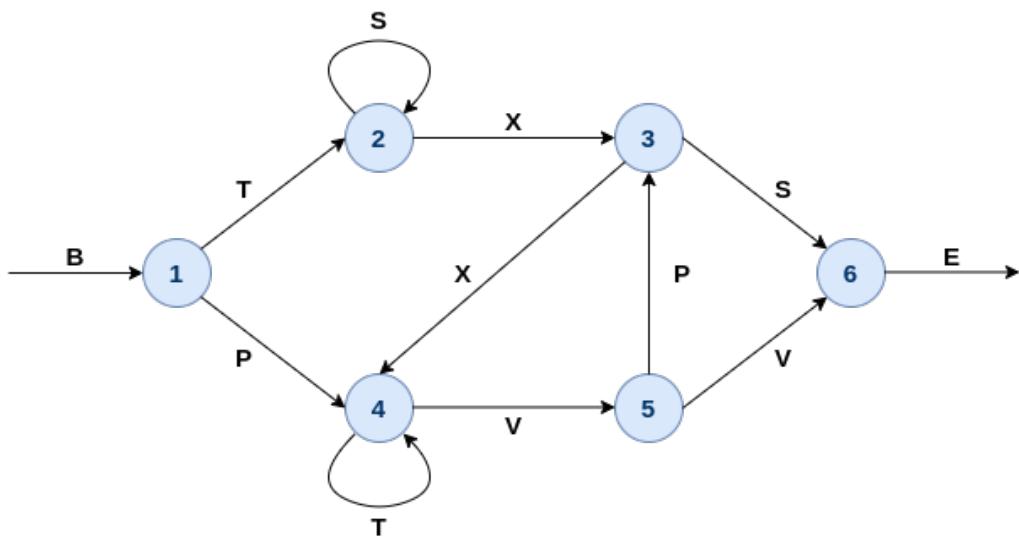


Figure 4.1: Flow diagram to generate Reber grammar sequences

As shown in the above graph, a sequence is a true Reber sequence if

1. it contains only these characters: *B, E, P, S, T, V, X*
2. it starts with a *B* and ends with an *E*, *always!*
3. it strictly follows the transition diagram shown in figure 4.1.

4 Dataset

After starting at B , we move from one node to the next until we reach E . If we have more than one path to choose from, we can randomly select one of them. Two of the allowed characters, S and T , have a self-loop at nodes 2 and 4, respectively, meaning we can have more than two consecutive S s and T s. For example, the following table shows some of the true and false Reber sequences from our dataset:

True Reber sequences	False Reber sequences
BPTVPXTTVVE	BTTVPXTVPSE
BTSSXXTTVVE	BPSXXTTVPSE
BTXXVPXTVVE	BPSSSXXTVVE
BPTTTTVPSE	BTTTVPXTVVE
	BPSSXXTTVVE

Table 4.1: A few examples of true and false Reber sequences.

Since, to our knowledge, there is no public dataset of Reber sequences available, we created our own dataset of 25000 true and false Reber sequences combined. We follow the same flow, as shown in figure 4.1, to generate true Reber sequences. Then, by slightly altering the same procedure, we generate false Reber sequences.

We then split this dataset into a train-test set, details of which are shown in the following bar chart:

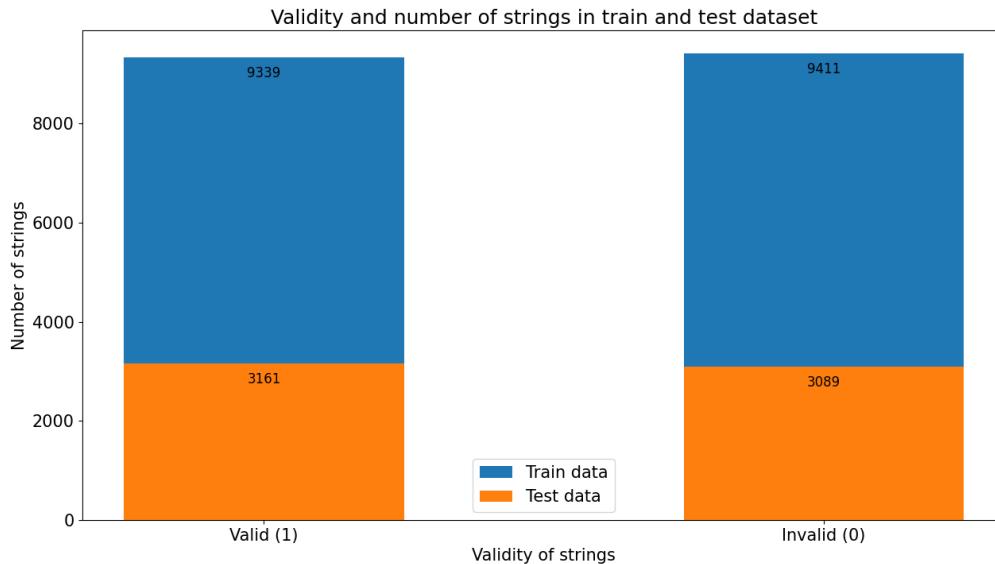


Figure 4.2: Number of valid (true) and invalid (false) Reber sequences in our train-test set

4 Dataset

Our dataset has 12500 true Reber sequences and 12500 false, thus totaling 25000. Although the true Reber sequences only contain allowed characters (i.e., *B*, *E*, *P*, *S*, *T*, *V*, *X*), the false sequences may contain any character from A to Z.

Out of 25000 total sequences, our train-test set contains 18750 and 6250 sequences, respectively. Furthermore, this train-test individually include the following number of true and false Reber sequences:

	Training set	Test set
Valid (true)	9339	3161
Invalid (false)	9411	3089

Table 4.2: The number of true and false Reber sequences in our train-test set.

Next, we visualize the string length distribution that shows not only the minimum and maximum string length but also the number of strings for each length value.

The following plot displays the string length distribution of our entire dataset:

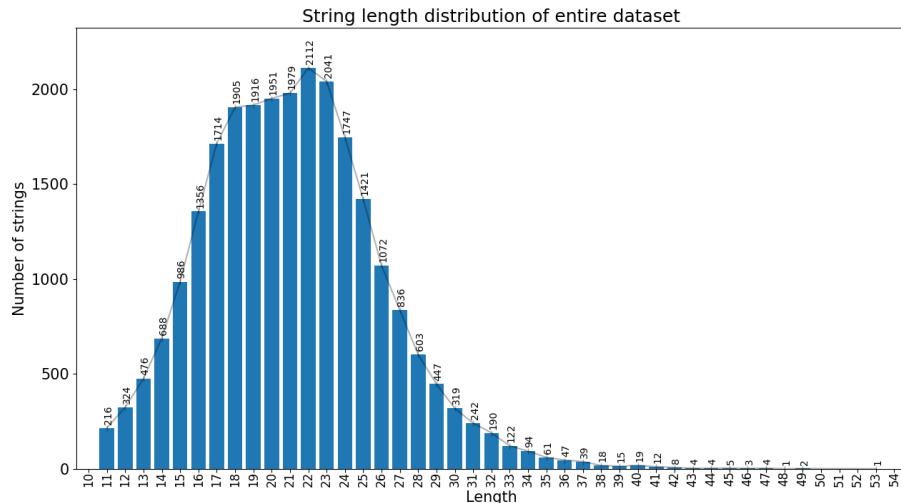


Figure 4.3: String length distribution of entire dataset

As we can see in the above plot, in our entire dataset, the minimum string length is 11 with 216 sequences, and the maximum string length is 53 with 1 sequence. The highest number of strings, 2112 strings, in our entire dataset is of length 22.

The following plot displays the string length distribution of our train dataset:

4 Dataset

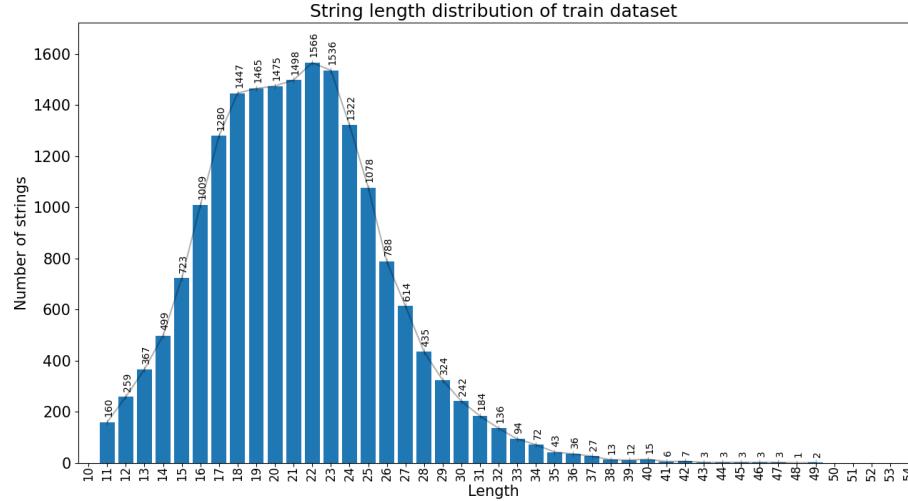


Figure 4.4: String length distribution of train dataset

As we can see in the above plot, in our train dataset, the minimum string length is 11 with 160 sequences, and the maximum string length is 49 with 2 sequence. The highest number of strings, 1566 strings, in our train dataset is of length 22.

The following plot displays the string length distribution of our test dataset:

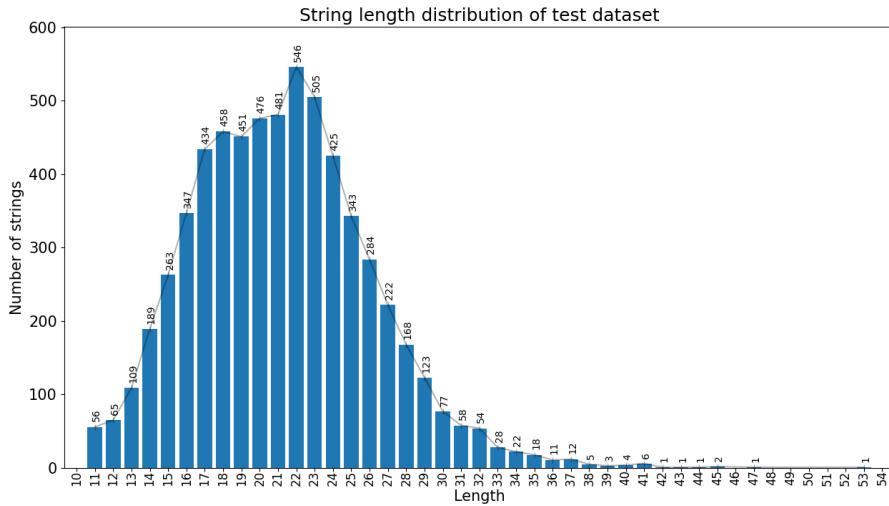


Figure 4.5: String length distribution of test dataset

As we can see in the above plot, in our test dataset, the minimum string length is 11 with 56 sequences, and the maximum string length is 53 with 1 sequence. The highest number of strings, 546 strings, in our test dataset is of length 22.

5 Sparse Recurrent Neural Networks

This chapter explains the experiments described in section 1.3, but before that, we first describe our base model, against which we will compare the performance of our pruning experiment.

5.1 Base model

Apart from the standard input and output layer, our base model is a regular RNN model followed by a linear layer and stacked upon an embedding layer, as shown in the following figure:

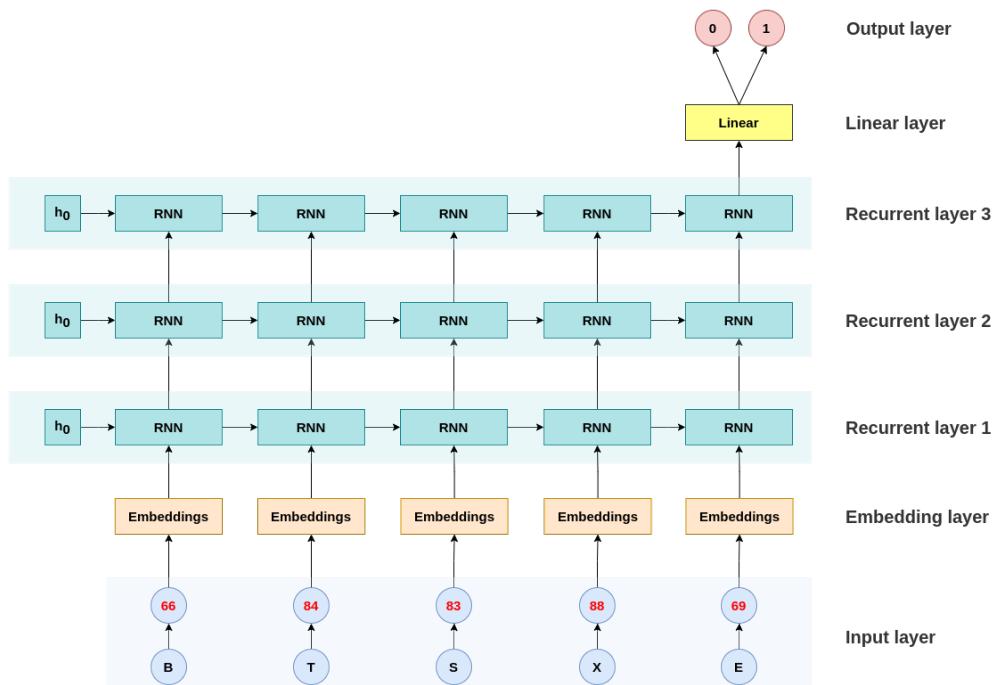


Figure 5.1: Visualizing training of a recurrent network with 3 recurrent layers, each layer consisting of 50 neurons.

Our base model starts with an input layer that feeds data to the model. Since our data is in text format, we convert each character into its corresponding ASCII value. This conversion is visualized in figure 5.1 with a valid example Reber sequence, **BTSXE**. Furthermore, we are working with batches, so all the sequences in a batch must be of equal length. Therefore, we apply zero-padding to ensure all the sequences in a given batch are of the same length.

After the input layer, the second layer is an embedding layer that transforms each input character into a fixed-length vector as visualized in figure 5.2. In our case, this embedding layer is of shape (128, 50), where 128 is the number of possible ASCII characters, and 50 is the size of each embedding vector known as embedding dimensions. The value of an embedding dimension must be equal to the number of neurons in the next layer.

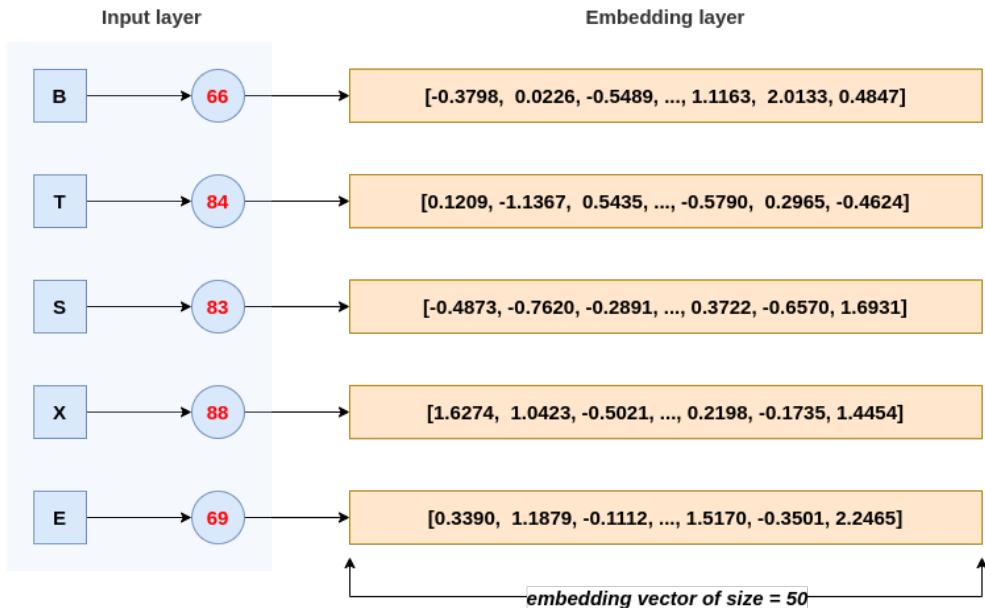


Figure 5.2: Generating embeddings of example Reber sequence after converting each character of this sequence into its corresponding ASCII value.

After the embedding layer, the subsequent three layers are recurrent, each layer containing 50 neurons. These layers start with an initial hidden state of shape (*batch size, hidden size*), comprising all zeros. In LSTM, these layers begin with one initial hidden state and one initial cell state, both of similar shape and comprising of all zeros. Each neuron in recurrent layers is an individual RNN unit with either Tanh or ReLU nonlinearity in case of vanilla recurrent network, LSTM unit in case of LSTM recurrent network and GRU unit in case of GRU recurrent network as visualized in the following

figure:

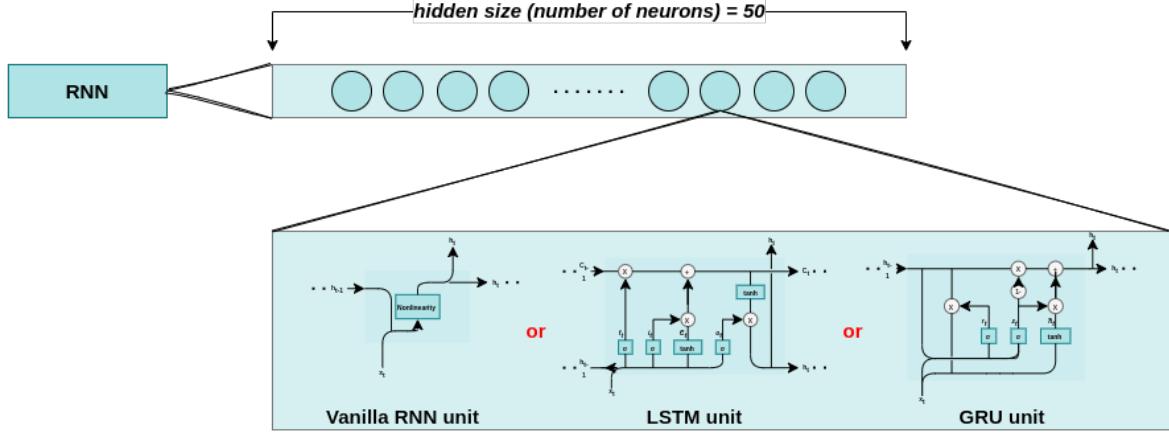


Figure 5.3: Visualizing a single recurrent layer from our model (fig. 5.1).
Each neuron is either an individual RNN unit (fig. 3.12) or an LSTM unit (fig. 3.14), or a GRU unit (fig. 3.15).

These recurrent layers are followed by a linear layer that takes input from the third recurrent layer and applies the following linear transformation on the input:

$$\hat{y} = xW^T + b \quad (5.1)$$

where \hat{y} is the output, W represents weight matrices of shape ($output_size = 2$, $input_size = 50$), x is the input, and b is bias of shape ($output_size = 2$).

Since the third recurrent layer's hidden size is 50, the linear layer's input is of size 50, and output is of size two because our target has only two classes, $\{0, 1\}$, as shown in figure 5.4.

For our example Reber sequence, the final output is in the form of a float array of size two, for instance, $[-0.8169, 1.3229]$. To transform it into 0 or 1, take the index of maximum value from this array. After getting an output, the model calculates the cross-entropy loss and optimizes it via backpropagation through time to a minimum value to achieve better performance.

We train this model on our training dataset of 18750 sequences for 50 epochs, and to check its performance, we evaluate it on the test dataset of 6250 sequences. This training and evaluation employ the hyper-parameters shown in the following table.

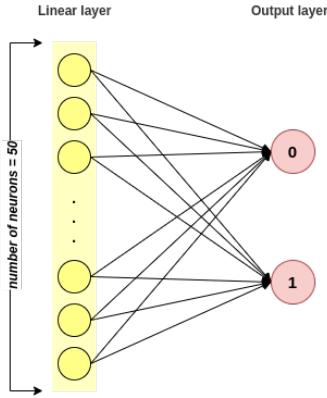


Figure 5.4: **Visualizing of linear layer's connection to output layer.**

Each of the 50 neurons from the linear layer is densely connected to both the output layer neurons.

Hyper-parameter	Value
Learning rate	0.001
Loss	Cross-Entropy Loss
Optimizer	Adam
Batch size	32

Table 5.1: Hyper-parameters used for training and evaluating the base model.

Since we work with four different RNN variants, namely RNN with Tanh nonlinearity, RNN with ReLU nonlinearity, LSTM, and GRU, we also perform this training four times.

As we have already established, training a recurrent network is computationally expensive and time-consuming. Therefore all four models in our experiment are trained on a Tesla K80 GPU.

5.2 Pruned Recurrent Neural Networks

One of the ways to introduce sparsity in recurrent networks is to prune weight below a certain threshold. As shown in figure 3.11, we have three different types of weights; input-to-hidden weights, hidden-to-hidden weights, and hidden-to-output weights. In our experiment, we individually and simultaneously prune input-to-hidden and hidden-to-hidden weights as envisioned in the following figure:

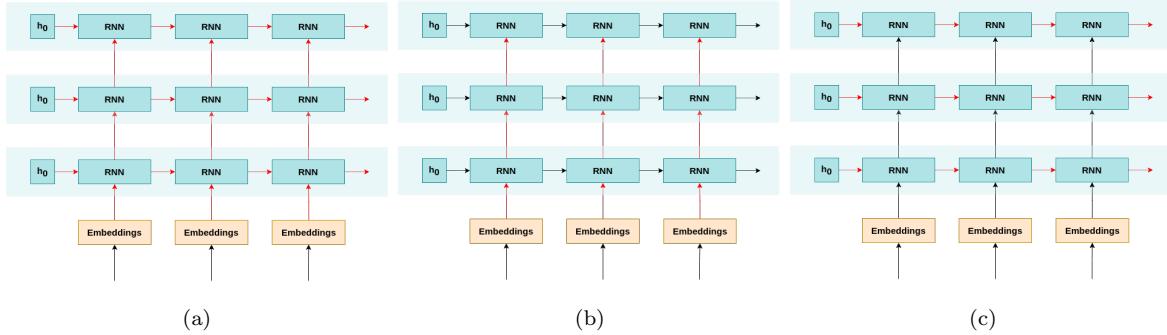


Figure 5.5: Simultaneously and individually pruning input-to-hidden and hidden-to-hidden weights. The red right arrow (\rightarrow) resembles the pruning of hidden-to-hidden weight, while the red up-arrow (\uparrow) resemble the pruning of input-to-hidden weight. Fig. 5.5a illustrates pruning both, input-to-hidden and hidden-to-hidden weights, fig. 5.5b illustrates pruning only input-to-hidden weights, and fig. 5.5c illustrates pruning only hidden-to-hidden weights.

The threshold based on which we prune weights is calculated based on the percent of weights to prune. Therefore, for example, to prune $p = 10\%$ of weights for a given layer, the threshold is the 10th percentile of all the **absolute** weights in that layer (code A.1). In our experiment, as shown in figure 1.2, we go from percent $p = 10$ to 100 while incrementing p by 10 after each round.

Using this threshold value, we create a binary tensor called a mask, where 1 corresponds to the **absolute** weight value above the threshold, and 0 corresponds to the **absolute** weight value below the threshold (code A.2) in the corresponding weight matrix. Element-wise multiplication of this binary mask with the corresponding weight matrix zero-outs the absolute weight values below the threshold value.

To understand this binary mask generation and mask multiplication, consider the following example weight matrix of shape $(3, 3)$ with float values between -1 and 1 :

$$\begin{bmatrix} -0.685 & 0.530 & -0.464 \\ -0.534 & 0.828 & 0.045 \\ -0.123 & 0.629 & -0.014 \end{bmatrix}$$

This weight matrix corresponds to a linear layer with 3 neurons densely connected to the next linear layer with 3 neurons as shown in the below figure:

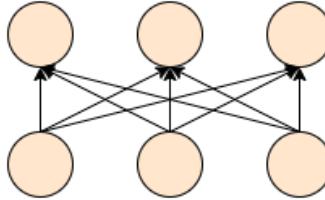


Figure 5.6: Two densely connected linear layers, each layer containing 3 neurons. This visualization corresponds to the example weight matrix shown above.

To prune 10% percent of weight values from the above matrix, we calculate the threshold by finding the 10th percentile of this matrix's **absolute** values, which is 0.0387. Using this threshold, we create a binary mask of shape $(3, 3)$, where values at positions that correspond to the below threshold **absolute** values in the weight matrix are kept 0 while other values are kept 1 as shown below:

$$\begin{bmatrix} -0.685 & 0.530 & -0.464 \\ -0.534 & 0.828 & 0.045 \\ -0.123 & 0.629 & \textcolor{red}{-0.014} \end{bmatrix} \longrightarrow \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & \textcolor{red}{0} \end{bmatrix}$$

By performing element-wise multiplication of this binary mask with our weight matrix, we get a pruned weight matrix as shown below:

$$\begin{bmatrix} -0.685 & 0.530 & -0.464 \\ -0.534 & 0.828 & 0.045 \\ -0.123 & 0.629 & \textcolor{red}{-0.014} \end{bmatrix} \circ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & \textcolor{red}{0} \end{bmatrix} = \begin{bmatrix} -0.685 & 0.530 & -0.464 \\ -0.534 & 0.828 & 0.045 \\ -0.123 & 0.629 & \textcolor{red}{0} \end{bmatrix}$$

The value 0 in this weight matrix represents the missing connections between the third neuron of the first and second linear layer. After applying the mask, the simple neural network shown in the figure 5.6 looks like below:

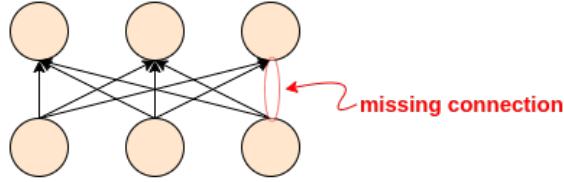


Figure 5.7: Two densely connected linear layers containing 3 neurons show the absence of connection between each layer’s third neuron after pruning. This visualization corresponds to the pruned weight matrix.

We follow a similar mask generation approach and element-wise multiplication for each corresponding weight matrix in each recurrent layer of our trained model. As shown above, this pruning process returns a pruned weight matrix for each corresponding weight matrix. After pruning, we evaluate pruned model’s performance on the test dataset to see how such pruning affects an already trained model’s overall accuracy.

Once we have the pruned model’s performance, the next step is to retrain the same pruned model to estimate the number of epochs required to regain the accuracy. We use the same pruned weight matrices throughout the entire retraining course. This helps us understand how long it takes for a recurrent network to relearn and update weight matrices to make up for the lost connections.

These pruning and retraining steps are repeated three times per RNN variant (i.e., RNN-Tanh, RNN-ReLU, LSTM, and GRU), first time for pruning both, input-to-hidden and hidden-to-hidden weights, the second time for pruning only input-to-hidden weights, and the third time for pruning only hidden-to-hidden weights.

5.3 Randomly Structured Recurrent Neural Networks

The pruning method to induce sparsity in a recurrent network works by canceling connections between neurons of two separate layers. While in this section, we explain another method to induce sparsity by generating random structures that are already sparse. This method was introduced for Artificial Neural Network by Stier et al. in [SG19], which we will revise to make it suitable for our goal.

The first step of this experiment is to generate random graphs irrespective of how they might correspond to a neural network. We generate 100 connected Watts–Strogatz (3.7.3) and Barabási–Albert (3.7.4) graphs using the graph generators provided by the networkx¹. Each of these graphs contains nodes between (10, 51), meaning the corresponding neural architecture will have (10, 51) neurons.

The networkx graph generators return undirected graphs. Therefore, the next is to make them directed. Converting a random graph into a Directed Acyclic Graph is vital to make it easier to compute layer indexing.

After converting a random graph into a DAG, the next step is to compute layer indexing of all the nodes in this DAG. This helps us identify which node of this DAG belongs to which layer in the resulting neural network. The recursive algorithm to compute layer indexing of a node (similar to the one used by Stier et al. in [SG19]) is as below:

Algorithm 2: Recursive function to compute layer indexing of nodes

Input: Graph G , Set of nodes V , Set of edges E

Output: Layer index mapping for each node $v \in V$

```

1 Def LayerIndexOf( $v$ ):
2    $l \leftarrow [-1]$                                 // List  $l$ , initialized with an element  $-1$ 
3   foreach  $s \in V | E_{s \rightarrow v}$  do
4     |  $l.append(LayerIndexOf(s))$ 
5   end
6   return ( $\max(l) + 1$ )
7  $indexes \leftarrow dict()$                       // Initializing an empty dictionary,  $indexes$ 
8 foreach  $v \in V$  do
9   |  $indexes[v] \leftarrow LayerIndexOf(v)$ 
10 end

```

¹NetworkX is an open-source Python library that empowers users to create and manipulate various graph structures.

This recursive function outputs layer indexing of all nodes in our DAG, based on which we generate our final neural architecture, which then we convert into a recurrent network by introducing hidden-to-hidden connections, i.e., recurrent connections.

Once we have our randomly structured recurrent network, the final step is to embed it between an embedding layer and a linear layer. The embedding layer transforms data from the input layer into a fixed-length vector and feeds them to our recurrent layer. The linear layer takes the last recurrent layer's output, applies a linear transformation shown in equation 5.1, and generates final outputs.

5.3.1 From Random Graph to recurrent network: An example

To begin, consider the following undirected graph containing 5 nodes:

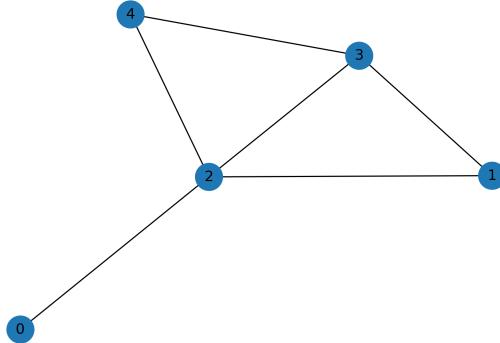


Figure 5.8: An undirected graph containing 5 nodes.

The next step is to convert this graph into a DAG. This conversion introduces directed edges from one node to another, as shown below:

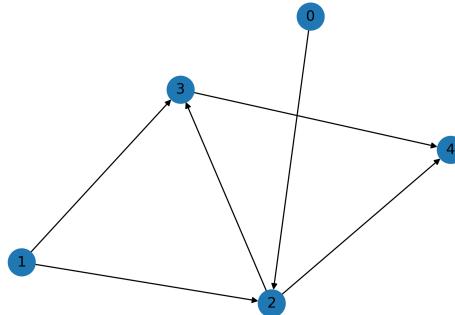


Figure 5.9: A directed graph containing 5 nodes.

The next step is to compute layer indexing of all 5 nodes in this DAG using algorithm 2. The following table shows the values of important variables in this algorithm and the final layer indexing:

Node v	Set of $s, \forall s \in V E_{s \rightarrow v}$	List l	Layer index of node v
0	{}	[-1]	0
1	{}	[-1]	0
2	{0, 1}	[-1, 0, 0]	1
3	{1, 2}	[-1, 0, 1]	2
4	{2, 3}	[-1, 1, 2]	3

Table 5.2: We first identify a set of all the parent nodes for a given node v , then append layer indexing of these parent nodes to the list l , and finally return $\max(l) + 1$ as layer index of node v .

The next step is to create a neural architecture using this layer indexing:

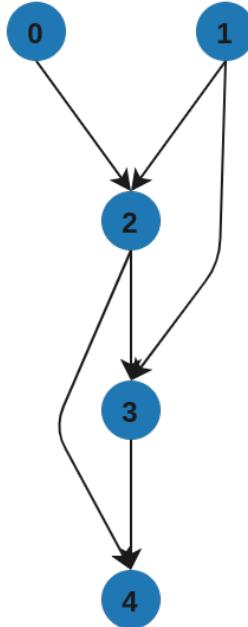


Figure 5.10: A neural architecture of 4 layers, generated using the layer indexing of nodes shown in table 5.2.

Once we have the neural architecture, as shown in figure 5.10, the next step is to introduce recurrent connections. This will give us our randomly structured recurrent network.

Attaching this randomly structured recurrent network between an embedding layer and a linear layer, as shown in the below figure, gives us our final model:

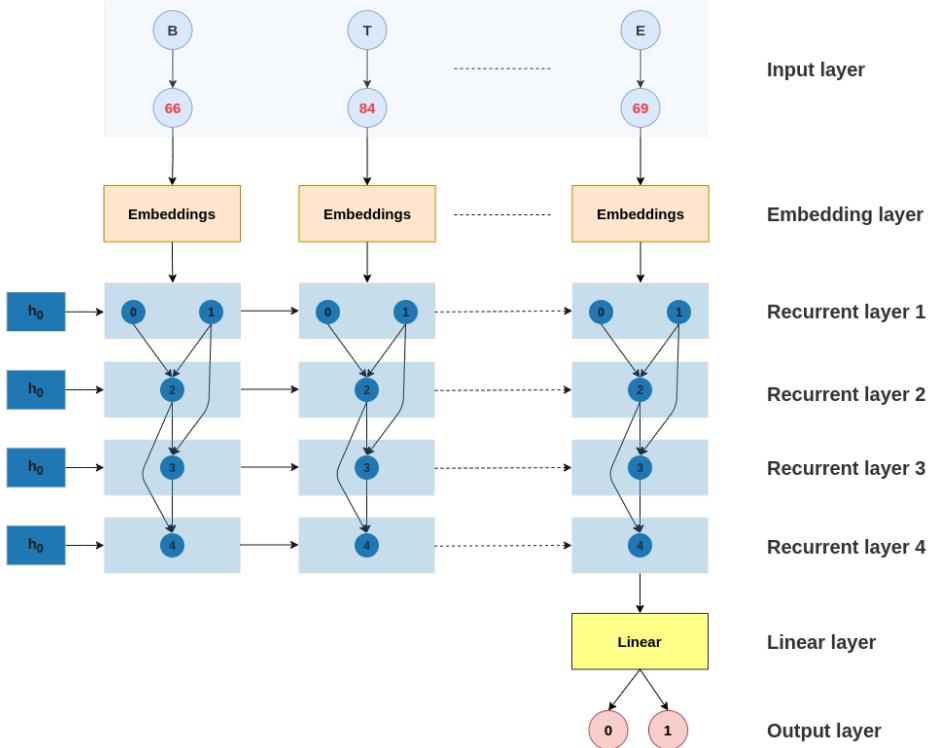


Figure 5.11: A complete trainable model with an input and output layer, an embedding layer, a linear layer, and a recurrent architecture generated from a random graph.

This is the final trainable model that we train on our training dataset for 15 epochs and evaluate on the test dataset.

Similarly, we generate 200 random graphs with nodes between (10, 51) for each RNN variant (i.e., RNN-Tanh, RNN-ReLU, LSTM, and GRU), convert them into a recurrent network, and create the full model for each randomly structured recurrent network, as shown in figure 5.11. We train them on our train dataset for 15 epochs with hyperparameters shown in table 5.1 except with batch size of 128. Finally, we evaluate them on our test dataset and store the performance that is useful in the next part of the experiment.

5.3.2 Identifying important graph properties

While training random structure recurrent networks, we also record specific graph properties (mentioned in section 3.7.2) of its base random graph. After training and evaluation are complete, we store its accuracy, creating a small dataset of 200 rows.

5 Sparse Recurrent Neural Networks

Since the graph properties in our small dataset are in different ranges, we apply MinMax scaler on our dataset to scale each graph property between a range (0, 1).

Next, we split this dataset into a train-test with a 0.9/0.1 ratio. We train three different regressor algorithms: Bayesian ridge regressor, Random Forest regressor, and Ada Boost regressor, on the train set with graph properties as features and its corresponding accuracy as the target.

Then, we evaluate each regressor algorithm on the test set and report an R-squared value to understand how our data fit each regressor model.

This chapter described our experiments with Sparse RNN in detail. After getting familiarized with our experiments, we present our experiments' results and their corresponding visualizations in the next chapter.

6 Experiment results

In this chapter, we present our results from experiments demonstrated previously in chapter 5. We begin with presenting base model performance, which we then compare with pruning results. Afterward, we show results of randomly structured recurrent networks and finally, results explaining the importance of graph properties of a base random graph to its corresponding recurrent network.

6.1 Base model performance

Our first plot visualizes the base model performance of RNN with Tanh nonlinearity. As we can see, except for a brief period of time initially, our model performs steadily at > 90% accuracy throughout the entire training phase.

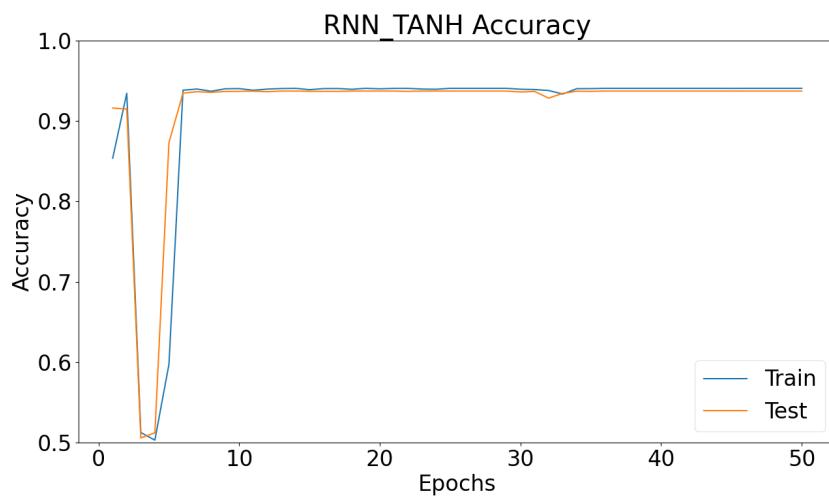


Figure 6.1: Base model performance of RNN with Tanh nonlinearity. This model is trained for 50 epochs with hyper-parameters shown in table 5.1.

6 Experiment results

Next is the base model performance of RNN with ReLU nonlinearity. As we can see in the following graph, our model is consistently above 96% accuracy throughout most of the training and evaluation phase.

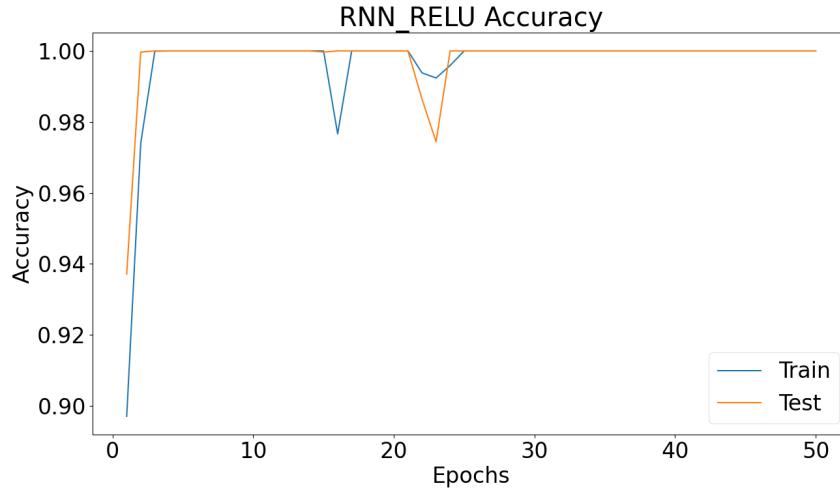


Figure 6.2: Base model performance of RNN with ReLU nonlinearity. This model is trained for 50 epochs with hyper-parameters shown in table 5.1.

Our base LSTM model's performance is consistent at 100% after beginning at around 93% in the first epoch, as shown in the following figure:

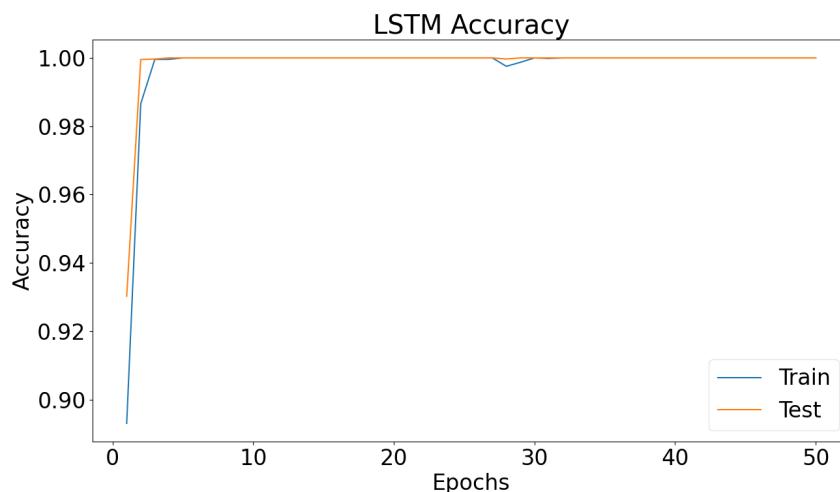


Figure 6.3: Base model performance of LSTM. This model is trained for 50 epochs with hyper-parameters shown in table 5.1.

Similar to LSTM, our base GRU model also performs consistently at 100% after beginning at around 93% in the first epoch, as shown in the following figure:

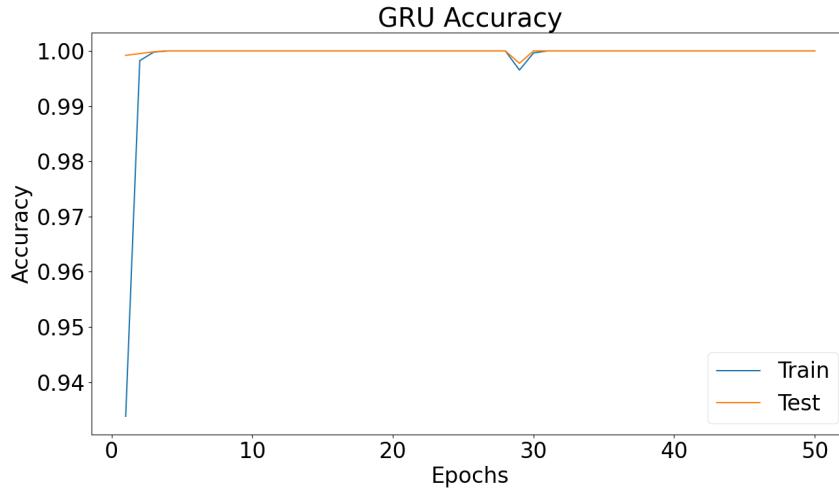


Figure 6.4: Base model performance of GRU. This model is trained for 50 epochs with hyper-parameters shown in table 5.1.

Next, we look at the performance of our base model after pruning it.

6.2 Base model performance after pruning

We begin with displaying the results of pruning both input-to-hidden and hidden-to-hidden weights for each RNN variant, then we proceed with the results of pruning only input-to-hidden weights, and later, the results of pruning only hidden-to-hidden weights.

In each section, we also present the visualizations showing the number of epochs required to regain accuracy after pruning each RNN variant's models.

6.2.1 Pruning input-to-hidden and hidden-to-hidden weights simultaneously

First, we look at performance of RNN with Tanh nonlinearity after applying 10% to 100% pruning. As we can see in figure 6.5, the model performs consistently at above 90% accuracy at 80% pruning, meaning the model retains near original performance with 80% fewer parameters. The model observes a sharp decline in performance at 90% pruning and goes to baseline accuracy at 100% pruning.

6 Experiment results

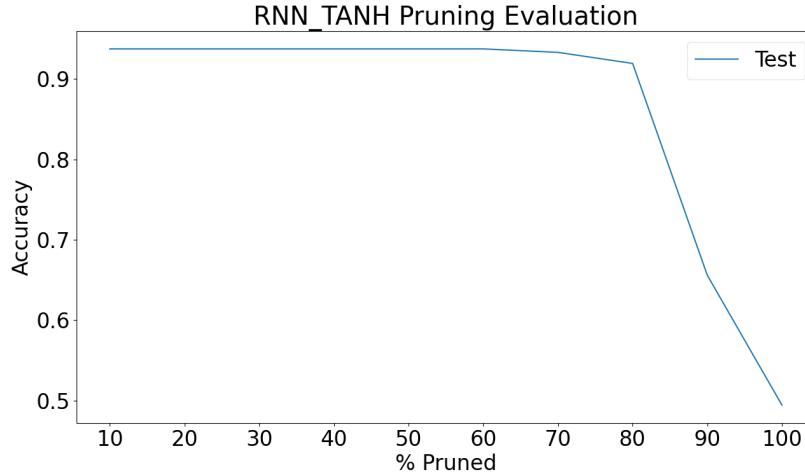


Figure 6.5: Base model performance of RNN_Tanh after pruning. The pruning starts from 10% and ends at 100% with an increment of 10 after each pruning round.

Next, we observe the number of epochs the model requires to regain the accuracy at 80%, 90%, and 100% pruning:

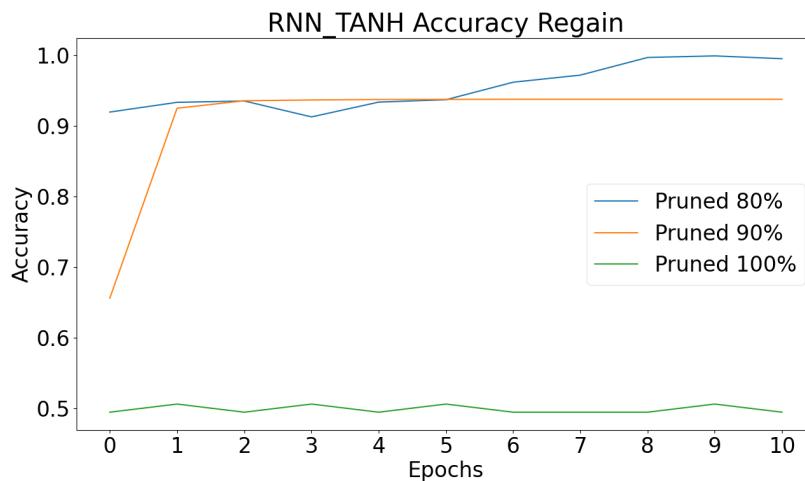


Figure 6.6: The number of epochs required to regain the accuracy of RNN_Tanh model after applying 80%, 90%, and 100% pruning.

As we can see in the above figure, after pruning 80% and 90% of input-to-hidden and hidden-to-hidden weights, the model regains the original above 90% performance only after one epoch, while the model never recovers at 100% pruning.

Next, we look at performance of RNN with ReLU nonlinearity after applying 10% to 100% pruning. As we can see in figure 6.7, the model performs consistently at above

6 Experiment results

90% accuracy at 70% pruning, meaning the model retains near original performance with 70% fewer parameters. The model observes a sharp decline in performance at 80% pruning and goes to baseline accuracy at 100% pruning.

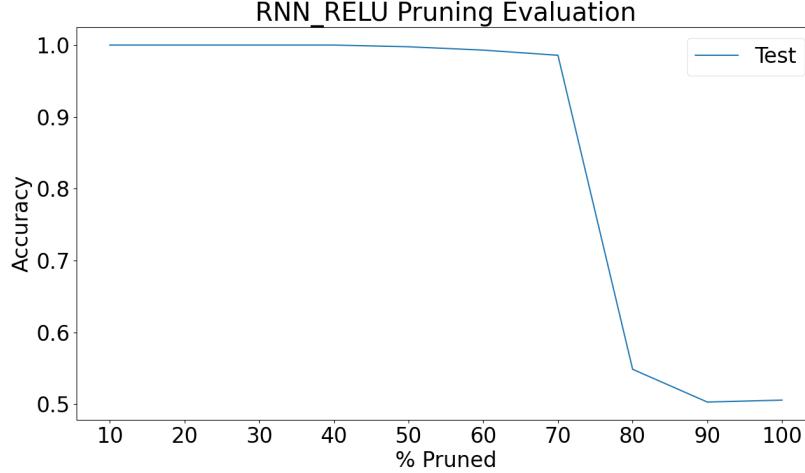


Figure 6.7: Base model performance of RNN_ReLU after pruning. The pruning starts from 10% and ends at 100% with an increment of 10 after each pruning round.

Next, we observe the number of epochs the model requires to regain the accuracy at 80%, 90%, and 100% pruning:

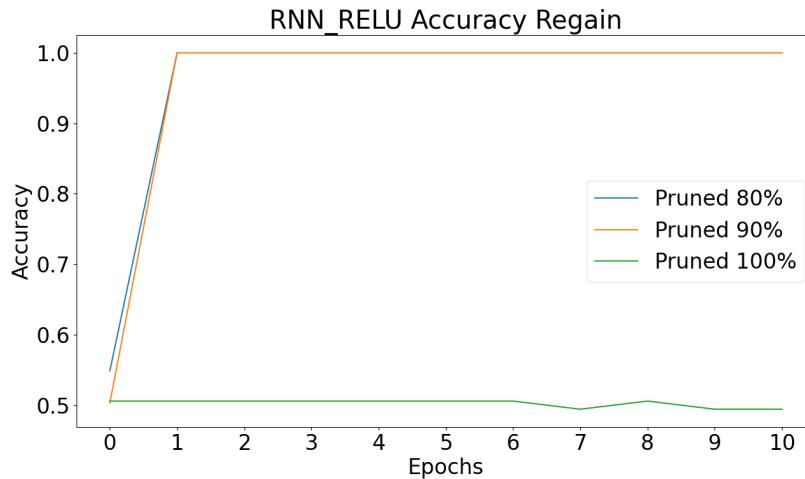


Figure 6.8: The number of epochs required to regain the accuracy of RNN_ReLU model after applying 80%, 90%, and 100% pruning.

Similar to RNN_Tanh, after pruning 80% and 90% of input-to-hidden and hidden-to-hidden weights, the model regains the original 100% accuracy only after one epoch, while the model never recovers at 100% pruning.

6 Experiment results

Next, we look at performance of LSTM model after applying 10% to 100% pruning. As we can see in figure 6.9, the model performs consistently at 100% accuracy at 60% pruning. The model observes a sharp decline in performance at 70% pruning and goes to baseline accuracy at 90% pruning.

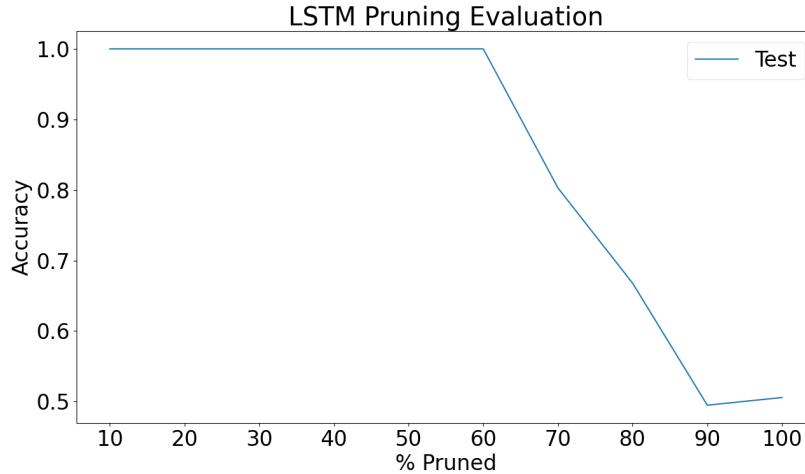


Figure 6.9: Base model performance of LSTM model after pruning. The pruning starts from 10% and ends at 100% with an increment of 10 after each pruning round.

Next, we observe the number of epochs the model requires to regain the accuracy at 70%, 80%, 90%, and 100% pruning:

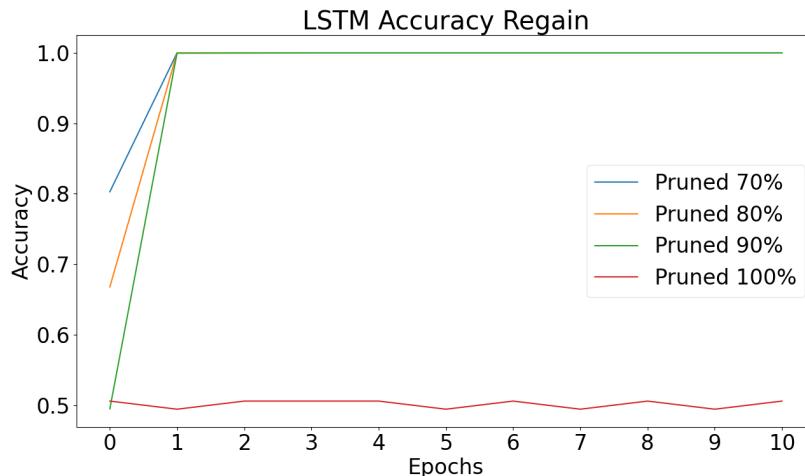


Figure 6.10: The number of epochs required to regain the accuracy of LSTM model after applying 70%, 80%, 90%, and 100% pruning.

Similar to RNN_Tanh and RNN_ReLU, after pruning 70%, 80% and 90% of input-to-hidden and hidden-to-hidden weights, the model regains the original 100% accuracy only

6 Experiment results

after one epoch, while the model never recovers at 100% pruning.

Finally, we look at performance of GRU model after applying 10% to 100% pruning. As we can see in figure 6.11, the model performs consistently at near 100% accuracy at 80% pruning. The model observes a sharp decline in performance at 90% pruning and goes to baseline accuracy at 100% pruning.

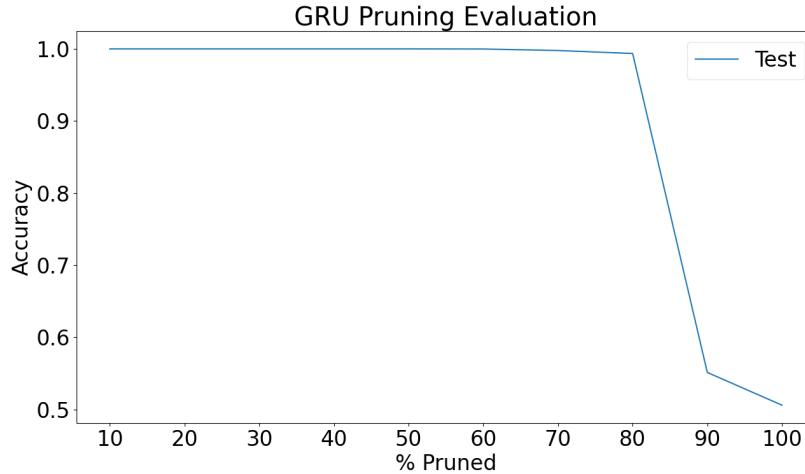


Figure 6.11: Base model performance of GRU model after pruning. The pruning starts from 10% and ends at 100% with an increment of 10 after each pruning round.

Next, we observe the number of epochs the model requires to regain the accuracy at 90%, and 100% pruning:

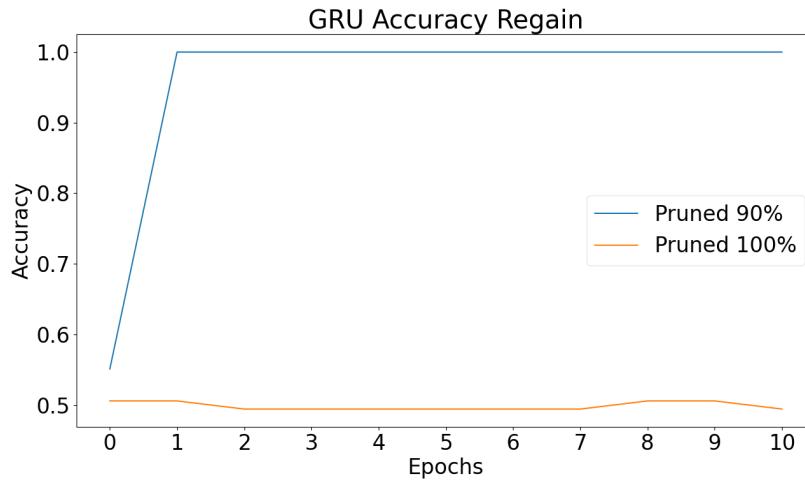


Figure 6.12: The number of epochs required to regain the accuracy of GRU model after applying 90%, and 100% pruning.

Similar to RNN_Tanh, RNN_ReLU, and LSTM, after pruning 90% of input-to-hidden and hidden-to-hidden weights, the model regains the original 100% accuracy only after one epoch, while the model never recovers at 100% pruning.

6.2.2 Pruning only input-to-hidden weights

In this section, we present the results of pruning only input-to-hidden weights in our base model.

First, we look at how pruning only input-to-hidden weights affect the performance of RNN with Tanh nonlinearity. As shown in figure 6.13, the model performs consistently at above 90% accuracy with 70% of input-to-hidden weights pruned. We observe a sharp drop in accuracy at 80% pruning.

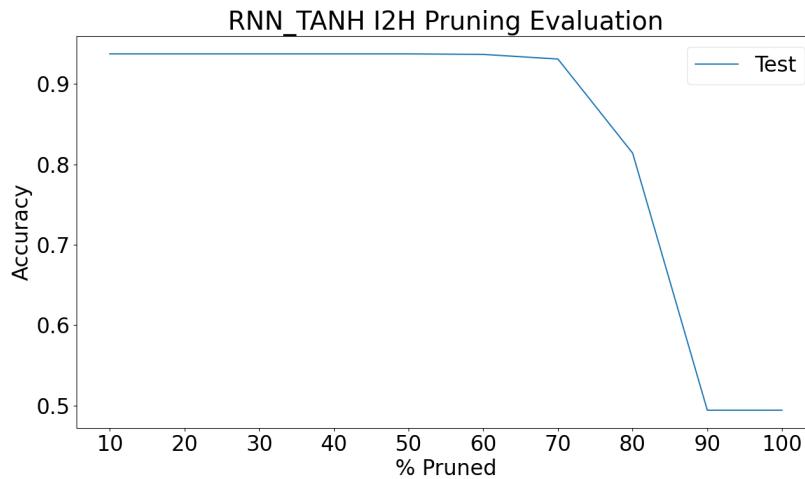


Figure 6.13: Base model performance of RNN with Tanh nonlinearity after pruning only input-to-hidden weights. The pruning starts from 10% and ends at 100% with an increment of 10 after each pruning round.

As shown in the above figure, the model starts performing worse from 80% pruning and above. Figure 6.14 shows the number of epochs it takes for this pruned model to regain accuracy after pruning 80%, 90%, and 100% of input-to-hidden weights.

As shown, the model regains above 90% accuracy only after one epoch with 80% pruning, while it takes two epochs to regain accuracy after 90% pruning. The model never recovers after pruning 100% of input-to-hidden weights.

6 Experiment results

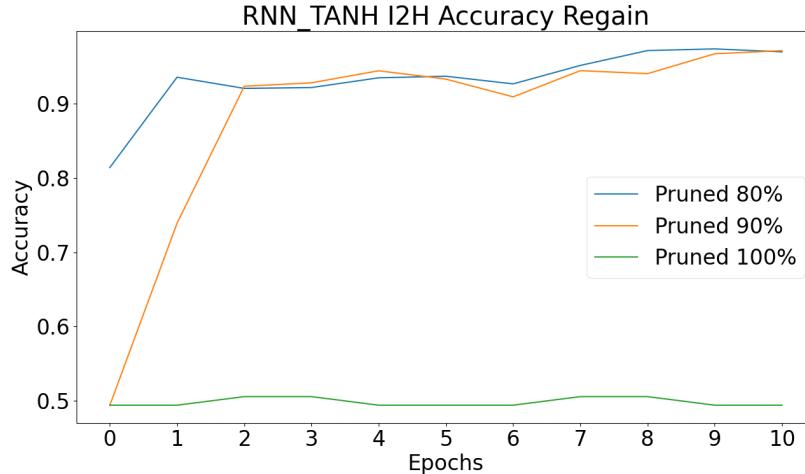


Figure 6.14: The number of epochs required to regain the accuracy of RNN_Tanh model after pruning 80%, 90%, and 100% of input-to-hidden weights.

Next, we look at how pruning only input-to-hidden weights affect the performance of RNN with ReLU nonlinearity. As shown in figure 6.15, the model performs consistently at above 90% accuracy with 80% of input-to-hidden weights pruned. We observe a sharp drop in accuracy at 90% pruning.

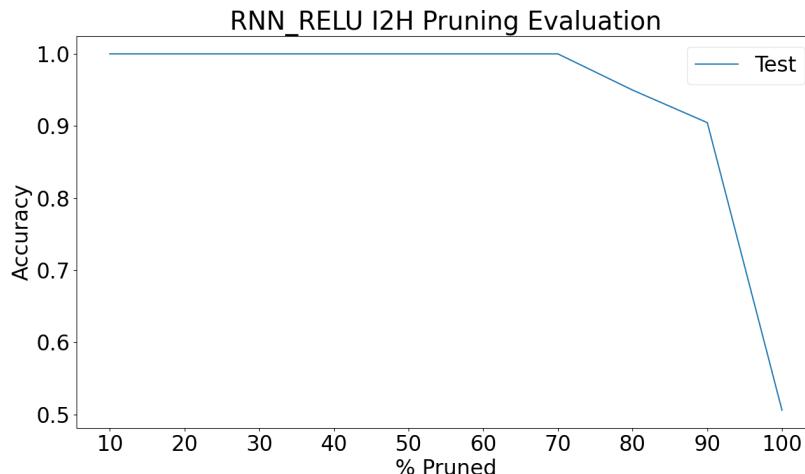


Figure 6.15: Base model performance of RNN with ReLU nonlinearity after pruning only input-to-hidden weights. The pruning starts from 10% and ends at 100% with an increment of 10 after each pruning round.

As shown in the above figure, the model starts dropping accuracy from 80% pruning and above. Figure 6.16 shows the number of epochs it takes for this pruned model to

6 Experiment results

regain accuracy after pruning 80%, 90%, and 100% of input-to-hidden weights.

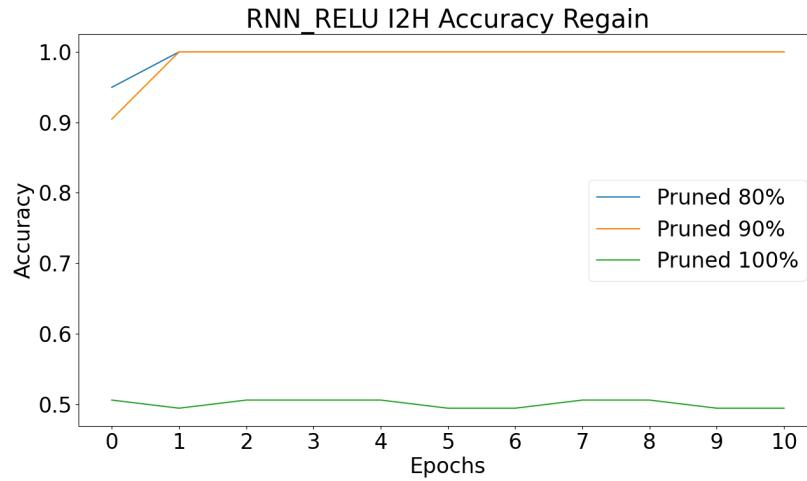


Figure 6.16: The number of epochs required to regain the accuracy of RNN_ReLU model after pruning 80%, 90%, and 100% of input-to-hidden weights.

As shown, the model regains above 90% accuracy only after one epoch with 80% and 90% pruning, while the model never recovers after pruning 100% of input-to-hidden weights.

Next, the following graph shows the effect of pruning only input-to-hidden weights on our LSTM model's performance:

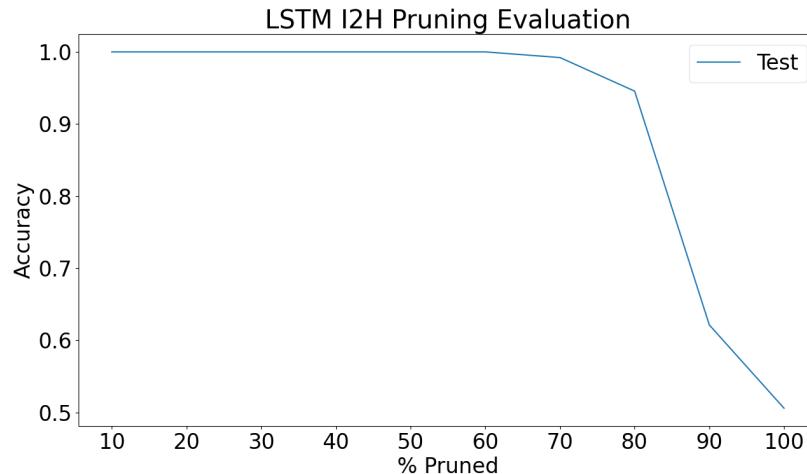


Figure 6.17: Base model performance of LSTM after pruning only input-to-hidden weights. The pruning starts from 10% and ends at 100% with an increment of 10 after each pruning round.

6 Experiment results

As we can see in the above graph, the model performs consistently at around 95% accuracy with 80% pruning. The model's performance drops sharply at 90% pruning and returns baseline accuracy at 100% pruning. Figure 6.18 shows the number of epochs it takes for the LSTM model to recover after pruning 80%, 90%, and 100% of its input-to-hidden weights.

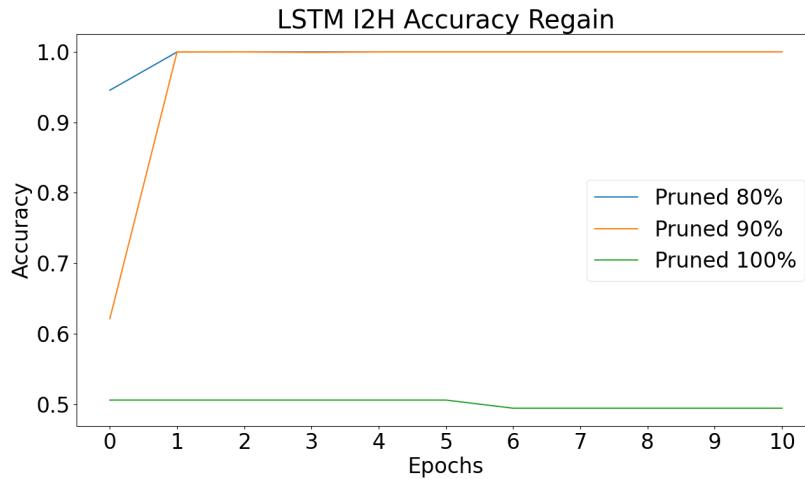


Figure 6.18: The number of epochs required to regain the accuracy of LSTM model after pruning 80%, 90%, and 100% of input-to-hidden weights.

Finally, the following graph shows the effect of pruning only input-to-hidden weights on our GRU model's performance:

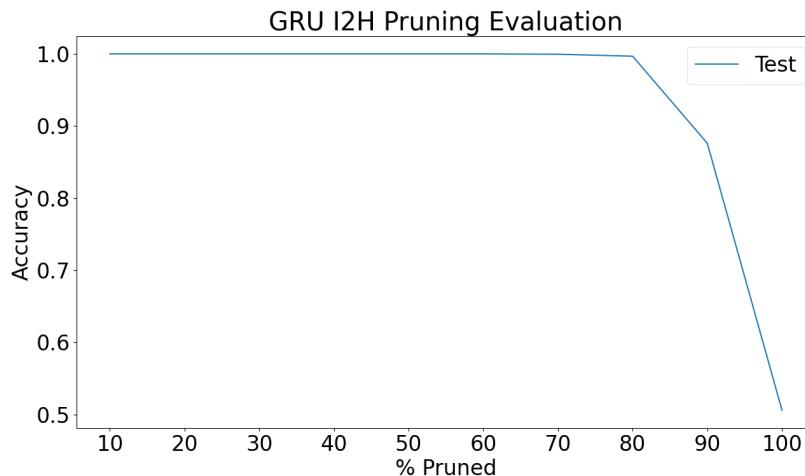


Figure 6.19: Base model performance of GRU after pruning only input-to-hidden weights. The pruning starts from 10% and ends at 100% with an increment of 10 after each pruning round.

6 Experiment results

As we can see in the above graph, the model performs consistently at above 95% accuracy with 80% pruning. The model's performance drops sharply at 90% pruning and returns baseline accuracy at 100% pruning. Figure 6.20 shows the number of epochs it takes for the GRU model to recover after pruning 90%, and 100% of its input-to-hidden weights.

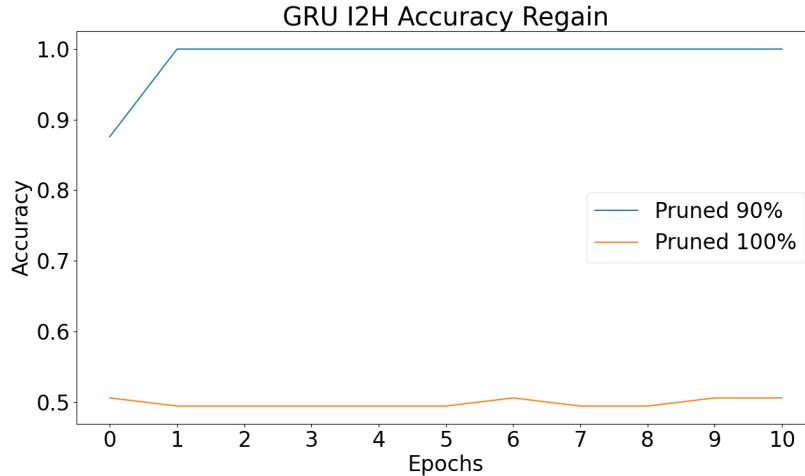


Figure 6.20: The number of epochs required to regain the accuracy of GRU model after pruning 90%, and 100% of input-to-hidden weights.

As we can see in the above graph, it only takes one epoch for the GRU model to recover from pruning 90% of its input-to-hidden weights. The model never recovers from pruning 100% of its input-to-hidden weights.

6.2.3 Pruning only hidden-to-hidden weights

In this section, we present the plots visualizing the effects of pruning only hidden-to-hidden weights on each RNN variant's performance.

Beginning with RNN with Tanh nonlinearity, as visualized in figure 6.21, the model performs consistently at above 90% at 80% of hidden-to-hidden weights pruned. We observe a sharp drop in accuracy at 90% pruning, while the model returns baseline accuracy at 100% pruning.

Since performance drops only at 90% and above of pruning, we perform an accuracy regain experiment for only 90% and 100% pruning of hidden-to-hidden weights as shown in figure 6.22.

6 Experiment results

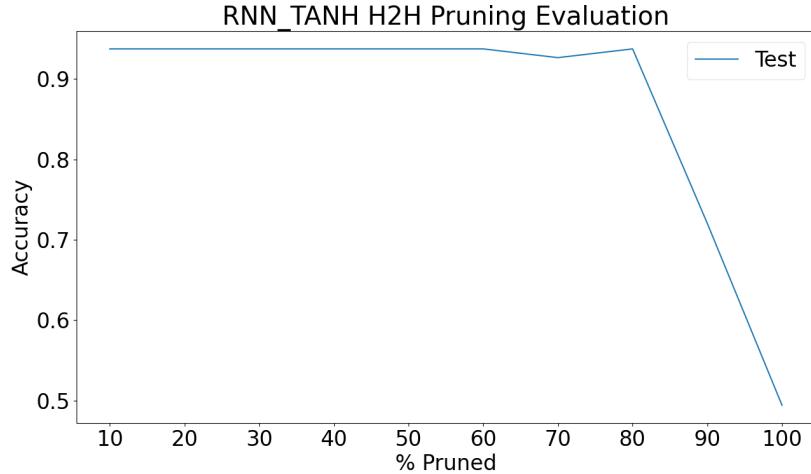


Figure 6.21: Base model performance of RNN with Tanh nonlinearity after pruning only hidden-to-hidden weights. The pruning starts from 10% and ends at 100% with an increment of 10 after each pruning round.

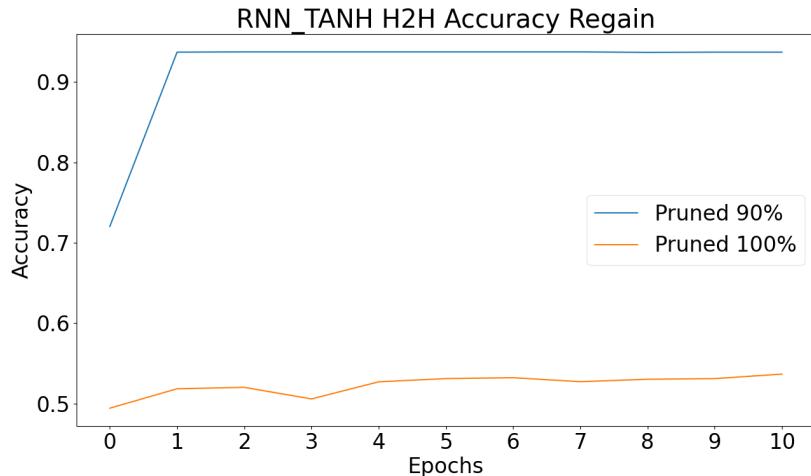


Figure 6.22: The number of epochs required to regain the accuracy of RNN_Tanh model after pruning 80%, 90%, and 100% of hidden-to-hidden weights.

As shown in the above graph, it only takes one epoch for the model to recover from 90% of hidden-to-hidden weights pruned, while the model never recovers after pruning 100% of hidden-to-hidden weights.

Next, we look at the effect of pruning hidden-to-hidden weights on the performance of RNN with ReLU nonlinearity. As shown in the below figure, the model performs

6 Experiment results

consistently at above 95% accuracy at 70% pruning and starts dropping the performance from 80% of pruning.

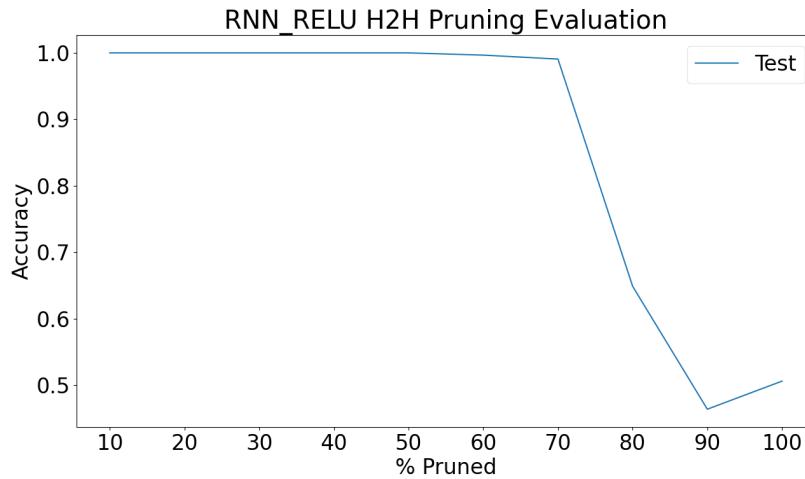


Figure 6.23: Base model performance of RNN with ReLU nonlinearity after pruning only hidden-to-hidden weights. The pruning starts from 10% and ends at 100% with an increment of 10 after each pruning round.

The following graph shows the required number of epochs for the RNN with ReLU nonlinearity to recover from 80%, 90%, and 100% pruning.

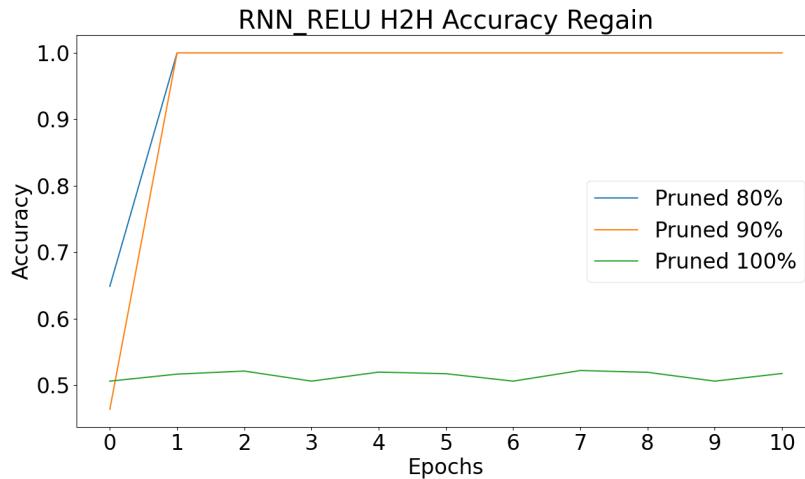


Figure 6.24: The number of epochs required to regain the accuracy of RNN_ReLU model after pruning 80%, 90%, and 100% of hidden-to-hidden weights.

As shown in the above figure, the model only requires one epoch to recover from 80%

6 Experiment results

and 90% of hidden-to-hidden weights pruning. The model, as before, never recovers from 100% pruning.

Next, we look at the effect of pruning hidden-to-hidden weights on the performance of LSTM, where, as shown in figure 6.25, the model performs consistently at around 95% accuracy at 90% pruning and drops the performance at 100% of pruning.

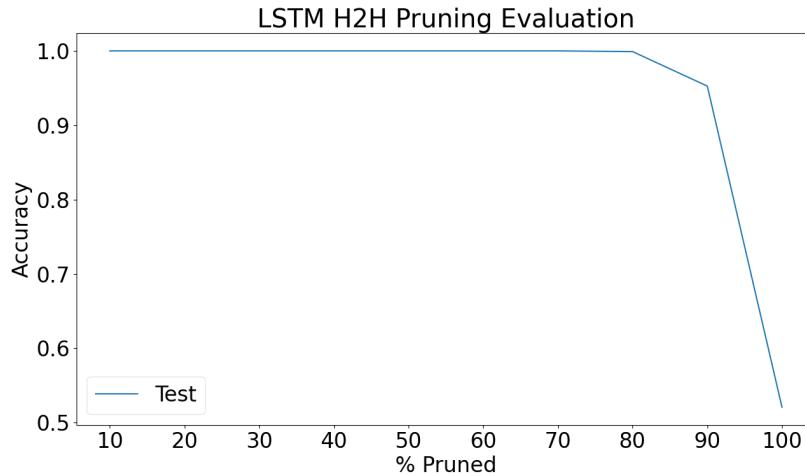


Figure 6.25: Base model performance of LSTM after pruning only hidden-to-hidden weights. The pruning starts from 10% and ends at 100% with an increment of 10 after each pruning round.

The following graph shows the required number of epochs for the LSTM to recover from 100% pruning.

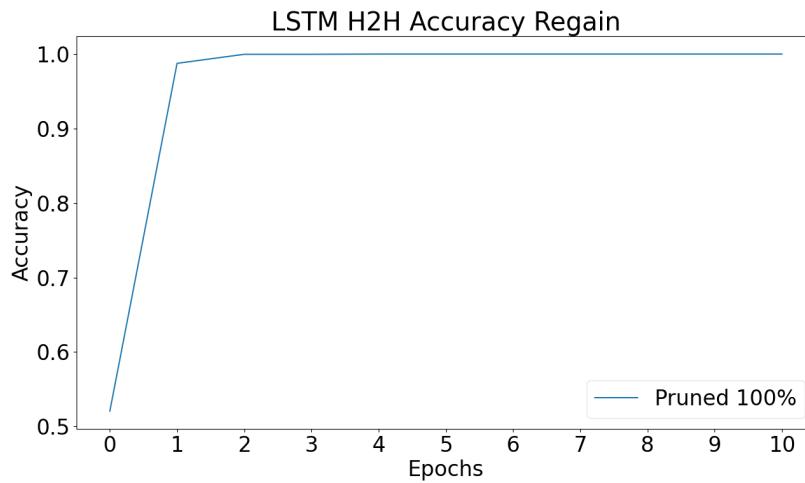


Figure 6.26: The number of epochs required to regain the accuracy of LSTM model after pruning 100% of hidden-to-hidden weights.

6 Experiment results

As shown in the above figure, even at 100% pruning of hidden-to-hidden weights, the model still recovers in just one epoch. The reason behind this recovery is the existence of a cell state, as explained in section 3.4.

Finally, similar to LSTM, GRU also performs consistently at around 95% at 90% of hidden-to-hidden weights pruning, as shown in the following figure:

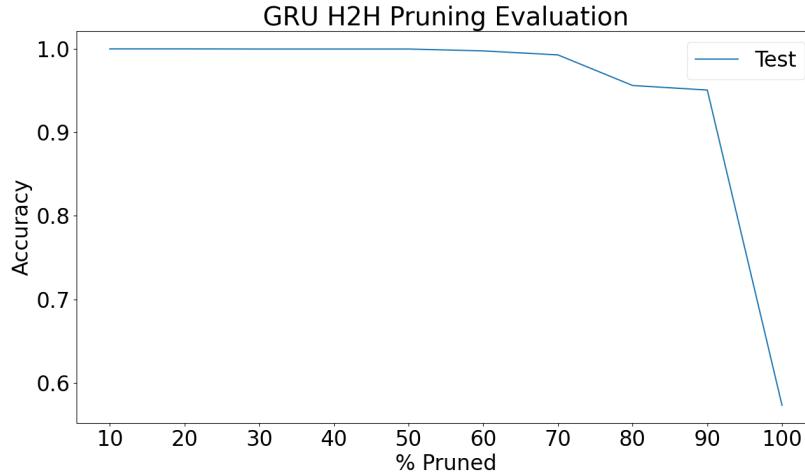


Figure 6.27: Base model performance of GRU after pruning only hidden-to-hidden weights. The pruning starts from 10% and ends at 100% with an increment of 10 after each pruning round.

The following graph shows the required number of epochs for the GRU to recover from 100% pruning.

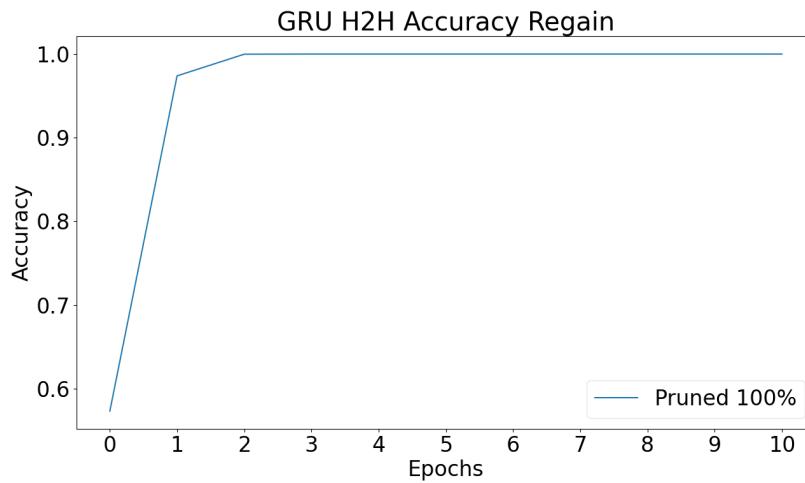


Figure 6.28: The number of epochs required to regain the accuracy of GRU model after pruning 100% of hidden-to-hidden weights.

Like the LSTM model, even at 100% pruning of hidden-to-hidden weights, the model still recovers in just one epoch, and after two epochs, the model returns 100% accuracy.

This concludes the results of the pruning experiment. In the next section, we present the results of our randomly structured recurrent networks.

6.3 Performance of Randomly Structured Recurrent Networks

Our second experiment was to construct a random structure recurrent network using a random graph and train it on the Reber grammar dataset. During this training, we record different graph properties of the base random graph and try to correlate it with the test performance. In this section, we present the results of this experiment.

6.3.1 Randomly Structured RNN with Tanh nonlinearity

We begin with the plot that compares the accuracy of the Watts–Strogatz model and Barabási–Albert model based random structure RNN_Tanh.

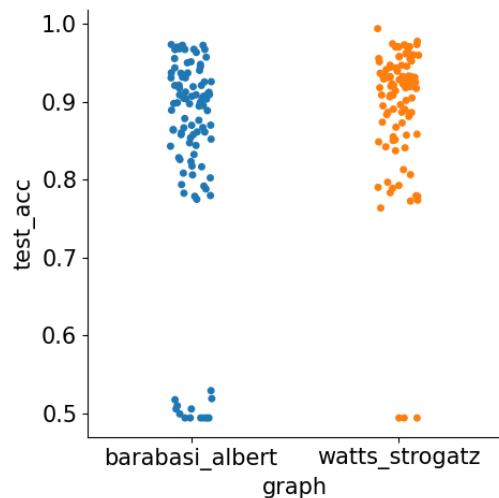


Figure 6.29: Plot comparing the performance of the Watts–Strogatz and Barabási–Albert model based random structure RNN_Tanh.

6 Experiment results

Next, the following table shows the Pearson correlation between test accuracy and different graph and recurrent network properties. The corresponding important plots are later shown in appendix C.1.

Property	Correlation with <i>test_acc</i>
layers	0.25
nodes	0.40
edges	0.38
source_nodes	0.35
diameter	-0.23
density	0.29
average_shortest_path_length	-0.27
eccentricity_var	-0.22
degree_var	-0.28
closeness_var	-0.46
nodes_betweenness_var	-0.49
edge_betweenness_var	-0.34

Table 6.1: Pearson correlation between test accuracy of RNN_Tanh and different graph and recurrent network properties

6.3.2 Randomly Structured RNN with ReLU nonlinearity

As with RNN_Tanh, we begin with the plot that compares the accuracy of the Watts–Strogatz model and Barabási–Albert model based random structure RNN_ReLU.

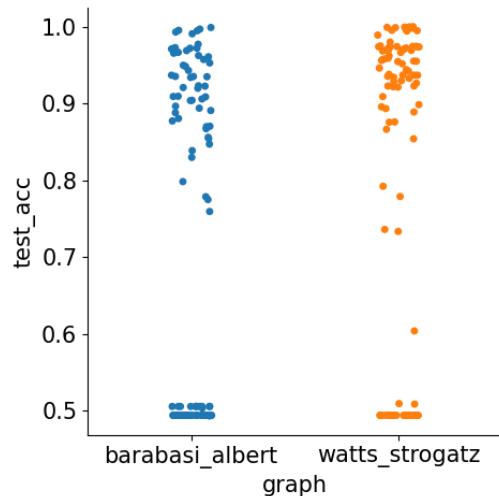


Figure 6.30: Plot comparing the performance of the Watts–Strogatz and Barabási–Albert model based random structure RNN_ReLU.

6 Experiment results

Next, the following table shows the Pearson correlation between test accuracy and different graph and recurrent network properties. The corresponding important plots are later shown in appendix C.2.

Property	Correlation with <i>test_acc</i>
layers	0.30
nodes	0.44
edges	0.43
source_nodes	0.47
diameter	-0.27
density	0.15
average_shortest_path_length	-0.25
eccentricity_var	-0.24
degree_var	-0.26
closeness_var	-0.39
nodes_betweenness_var	-0.41
edge_betweenness_var	-0.30

Table 6.2: Pearson correlation between test accuracy of RNN_ReLU and different graph and recurrent network properties

6.3.3 Randomly Structured LSTM

As before, we begin with the plot that compares the accuracy of the Watts–Strogatz model and Barabási–Albert model based random structure LSTM.

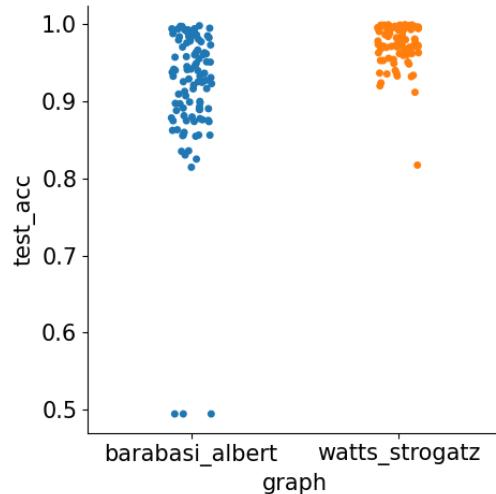


Figure 6.31: Plot comparing the performance of the Watts–Strogatz and Barabási–Albert model based random structure LSTM.

6 Experiment results

Next, the following table shows the Pearson correlation between test accuracy and different graph and recurrent network properties. The corresponding important plots are later shown in appendix C.3.

Property	Correlation with <i>test_acc</i>
layers	0.28
nodes	0.44
edges	0.42
source_nodes	0.57
diameter	-0.32
density	0.29
average_shortest_path_length	-0.36
eccentricity_var	-0.30
degree_var	-0.39
closeness_var	-0.51
nodes_betweenness_var	-0.56
edge_betweenness_var	-0.44

Table 6.3: Pearson correlation between test accuracy of LSTM and different graph and recurrent network properties

6.3.4 Randomly Structured GRU

As with LSTM, we begin with the plot that compares the accuracy of the Watts–Strogatz model and Barabási–Albert model based random structure GRU.

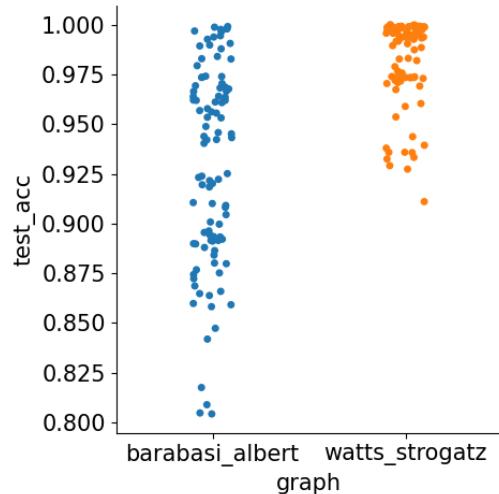


Figure 6.32: Plot comparing the performance of the Watts–Strogatz and Barabási–Albert model based random structure GRU.

6 Experiment results

Next, the following table shows the Pearson correlation between test accuracy and different graph and recurrent network properties. The corresponding important plots are later shown in appendix C.4.

Property	Correlation with <i>test_acc</i>
layers	0.34
nodes	0.49
edges	0.49
source_nodes	0.74
diameter	-0.20
density	0.34
average_shortest_path_length	-0.23
eccentricity_var	-0.21
degree_var	-0.58
closeness_var	-0.67
nodes_betweenness_var	-0.52
edge_betweenness_var	-0.26

Table 6.4: Pearson correlation between test accuracy of GRU and different graph and recurrent network properties

Positive Pearson correlation values in tables 6.1, 6.2, 6.3, and 6.4 indicate that an increase in the property variable increases test accuracy. Similarly, a negative Pearson correlation in the same tables indicates that a decrease in the property variable decreases the test accuracy. A higher positive or negative correlation indicates a strong relationship between a graph property and test corresponding test accuracy.

6.4 Performance prediction

While training randomly structured recurrent networks, we create a small dataset by storing all the graph and recurrent network properties and their corresponding test accuracy. We train three regressor algorithms on this dataset, and based on the resulting R-squared value, we determine if it is possible to predict a randomly structured recurrent network's performance.

For each RNN variant, the following table shows the R-squared values from each regressor algorithm we used:

6 Experiment results

RNN Variant	Bayesian Ridge	Random Forest	AdaBoost
RNN_Tanh	0.47919	0.43163	0.35698
RNN_ReLU	0.36075	0.61504	0.53469
LSTM	0.37206	0.57933	0.59514
GRU	0.67224	0.87635	0.78313

Table 6.5: R-squared values from each regressor algorithm for each RNN variant help determine our data's closeness to the fitted regression line. All four RNN variants presented in this table are randomly structured.

In simple words, the R-squared value helps us to understand how close our data is to the fitted regression line. A higher R-square value usually means that our data is a better fit for the model.

In our case, RNN_Tanh has below 0.5 R-squared value for all three regressors, while RNN_ReLU has a good R-squared value with Random Forest regressor (0.61). In contrast to RNN_ReLU, LSTM has a better R-squared value with AdaBoost regressor, while GRU has excellent R-squared values for all three regressor algorithms.

Based on these results, we further experimented with Random Forest regressor and data from randomly structured RNN_ReLU and randomly structured GRU to extract feature importance scores of different properties under different circumstances. Results from this experiment are shown in appendix [B](#).

7 Discussion

In this section, we discuss our findings from each experiment. Beginning from base model performance, we review the results of Sparse RNNs.

7.1 Base model performance

For our experiment purpose, we developed custom recurrent models using PyTorch, such that we can easily modify the weights based on our requirements. Each of our recurrent models, i.e., RNN with Tanh nonlinearity, RNN with ReLU nonlinearity, LSTM, and GRU, performs consistently with over 90% accuracy after training for only 50 epochs, as shown in section [6.1](#).

We used the same models to perform pruning experiments, a technique to generate sparsity in recurrent networks.

7.2 Pruning recurrent networks

This experiment helped answer the first three pruning-related research questions given in section [1.2](#).

Our pruning experiment was divided into three separate sub-experiments: pruning input-to-hidden and hidden-to-hidden weights simultaneously, pruning only input-to-hidden weights and pruning only hidden-to-hidden weights. For each sub-experiment, we also find the number of epochs required to regain the original performance.

While pruning both types of weights simultaneously, we found we can safely prune 80% of RNN_Tanh, 70% of RNN_ReLU, 60% of LSTM, and 80% of GRU. Afterward, we

retrained these pruned models to find that for each RNN variant, we require only one epoch to regain the original performance, while these models never recover after pruning 100% of weights.

While pruning only input-to-hidden weights, we found we can safely prune 70% of RNN_Tanh, 70% of RNN_ReLU, 70% of LSTM, and 80% of GRU. Afterward, we retrained these pruned models to find that for each RNN variant, we mostly require only one epoch to regain the original performance and just two epochs in the case of RNN_Tanh with 80% pruning. These models never recover after pruning 100% of the weights.

While pruning only hidden-to-hidden weights, we found we can safely prune 80% of RNN_Tanh, 70% of RNN_ReLU, 90% of LSTM, and 90% of GRU. Afterward, we retrained these pruned models to find that for each RNN variant, we require only one epoch to regain the original performance. RNN_Tanh and RNN_ReLU models never recover after pruning 100% of the weights, while LSTM and GRU still regain the original performance even with 100% pruning of hidden-to-hidden weights.

7.3 Randomly structured recurrent networks

This experiment helped answer the remaining two research questions given in section 1.2.

The resulting Pearson correlation values from training randomly structured recurrent networks help identify important graph and recurrent network properties for each RNN variant.

Based on this correlation, we found closeness_var, nodes_betweenness_var and the number of nodes to be essential properties for randomly structured RNN_Tanh. For randomly structured RNN_ReLU, the essential properties are the number of nodes, the number of edges, the number of source nodes, and nodes_betweenness_var.

In the case of randomly structured LSTM, we found six essential properties, i.e., the number of nodes, the number of edges, the number of source nodes, closeness_var, nodes_betweenness_var, and edge_betweenness_var. Similarly, we found six essential

7 Discussion

properties for randomly structured GRU, namely, the number of nodes, the number of edges, the number of source nodes, degree_var, closeness_var, and nodes_betweenness_var.

Next, we trained three regressor algorithms to find if we can use graph properties of a base random graph for performance prediction based on how well our data fit a model. Based on the results, we found that RNN_Tanh has below 0.5 R-squared value for each regressor, meaning our data is a weak fit to all three regression models.

In the case of RNN_ReLU, we have a 0.61 R-squared value with Random Forest regressor, meaning our data is slightly moderate fit to this particular regressor. For LSTM, our model is again slightly moderate fit to AdaBoost regressor with 0.60 R-squared value.

GRU has the best R-squared value of 0.81 with Random Forest regressor, indicating our data from randomly structured GRU is a strong fit to the Random Forest regressor model.

8 Conclusion

The main objective of our thesis, as described in chapter 1, was to investigate the effects of sparsity in recurrent neural networks. In chapter 5, we thoroughly explained the methodology followed, with results presented in chapter 6. In this chapter, we briefly summarize our outcomes.

We followed two different methods to induce sparsity in recurrent networks. One method introduced sparsity by pruning a certain percent of input-to-hidden and hidden-to-hidden weights, and another method is to generate sparse structures based on random graphs.

Based on the results of pruning experiments, we conclude that it is possible to reduce the weight complexity of different RNN variants by more than 60%. Our experiments also confirmed that it is possible to regain the accuracy of a pruned model with only one epoch in most cases. One big difference between RNN_Tanh, RNN_ReLU, and LSTM, GRU is that LSTM and GRU can regain the original performance even with 100% hidden-to-hidden weight pruning.

Random structure experiments helped to identify the essential graph properties of a base random graph. In all four RNN variants we experimented with, two mutual graph properties are the number of nodes and node betweenness. The two most mutual and important graph properties are the number of edges and the number of source nodes.

The performance prediction experiment results identified that data from RNN_Tanh has a weak fit with all three regressors used. Data from RNN_ReLU and GRU has moderate and strong fit Random Forest regressor, respectively. Finally, LSTM has a moderate fit with AdaBoost regressor.

A Code

Code A.1: Calculating threshold for pruning

```
1 weights = []
2 for param, data in self.named_parameters():
3     if 'bias' not in param and key in param:
4         weights += list(data.cpu().data.abs().numpy().flatten())
5 threshold = np.percentile(np.array(weights), percent)
```

Code A.2: Calculating binary mask

```
1 masks = {}
2 for l, layer in enumerate(self.recurrent_layers):
3     masks[l] = []
4     for param, data in layer.named_parameters():
5         if 'bias' not in param and key in param:
6             mask = torch.ones(data.shape,
7                               dtype=torch.bool,
8                               device=data.device)
9             mask[torch.where(abs(data) < threshold)] = False
10            masks[l].append(mask)
```

B Tables

	All properties	Only nodes and edges	Without nodes and edges	Only variances
Random Forest R^2	0.61	0.71	-0.05	0.06
Property	Feature importance			
layers	0.04		0.06	
nodes	0.02	0.23		
edges	0.01	0.29		
source_nodes	0.08	0.29		
sink_nodes	0.11	0.19		
diameter	0.0		0.0	
density	0.03		0.03	
average_shortest_path_length	0.03		0.03	
eccentricity_mean	0.03		0.04	
eccentricity_var	0.04		0.06	0.21
eccentricity_std	0.05		0.09	
degree_mean	0.03		0.03	
degree_var	0.03		0.06	0.17
degree_std	0.03		0.04	
closeness_mean	0.04		0.05	
closeness_var	0.07		0.09	0.22
closeness_std	0.07		0.08	
nodes_betweenness_mean	0.05		0.08	
nodes_betweenness_var	0.08		0.08	0.24
nodes_betweenness_std	0.07		0.07	
edge_betweenness_mean	0.05		0.06	
edge_betweenness_var	0.02		0.03	0.16
edge_betweenness_std	0.02		0.02	

Table B.1: Structural properties of the randomly structured RNN_ReLU's base random graphs and their feature importance scores for the Random Forest regressor under different circumstances.

	All properties	Only nodes and edges	Without nodes and edges	Only variances
Random Forest R^2	0.87	0.75	0.78	0.76
Property	Feature importance			
layers	0.01		0.04	
nodes	0.01	0.11		
edges	0.01	0.12		
source_nodes	0.67	0.71		
sink_nodes	0.01	0.06		
diameter	0.0		0.01	
density	0.01		0.02	
average_shortest_path_length	0.01		0.02	
eccentricity_mean	0.01		0.02	
eccentricity_var	0.02		0.03	0.17
eccentricity_std	0.02		0.03	
degree_mean	0.01		0.03	
degree_var	0.01		0.03	0.15
degree_std	0.01		0.03	
closeness_mean	0.01		0.03	
closeness_var	0.02		0.11	0.36
closeness_std	0.01		0.16	
nodes_betweenness_mean	0.04		0.11	
nodes_betweenness_var	0.01		0.15	0.25
nodes_betweenness_std	0.02		0.14	
edge_betweenness_mean	0.03		0.03	
edge_betweenness_var	0.01		0.02	0.08
edge_betweenness_std	0.01		0.02	

Table B.2: Structural properties of the randomly structured GRU's base random graphs and their feature importance scores for the Random Forest regressor under different circumstances.

C Plots

C.1 Randomly structured RNN_Tanh

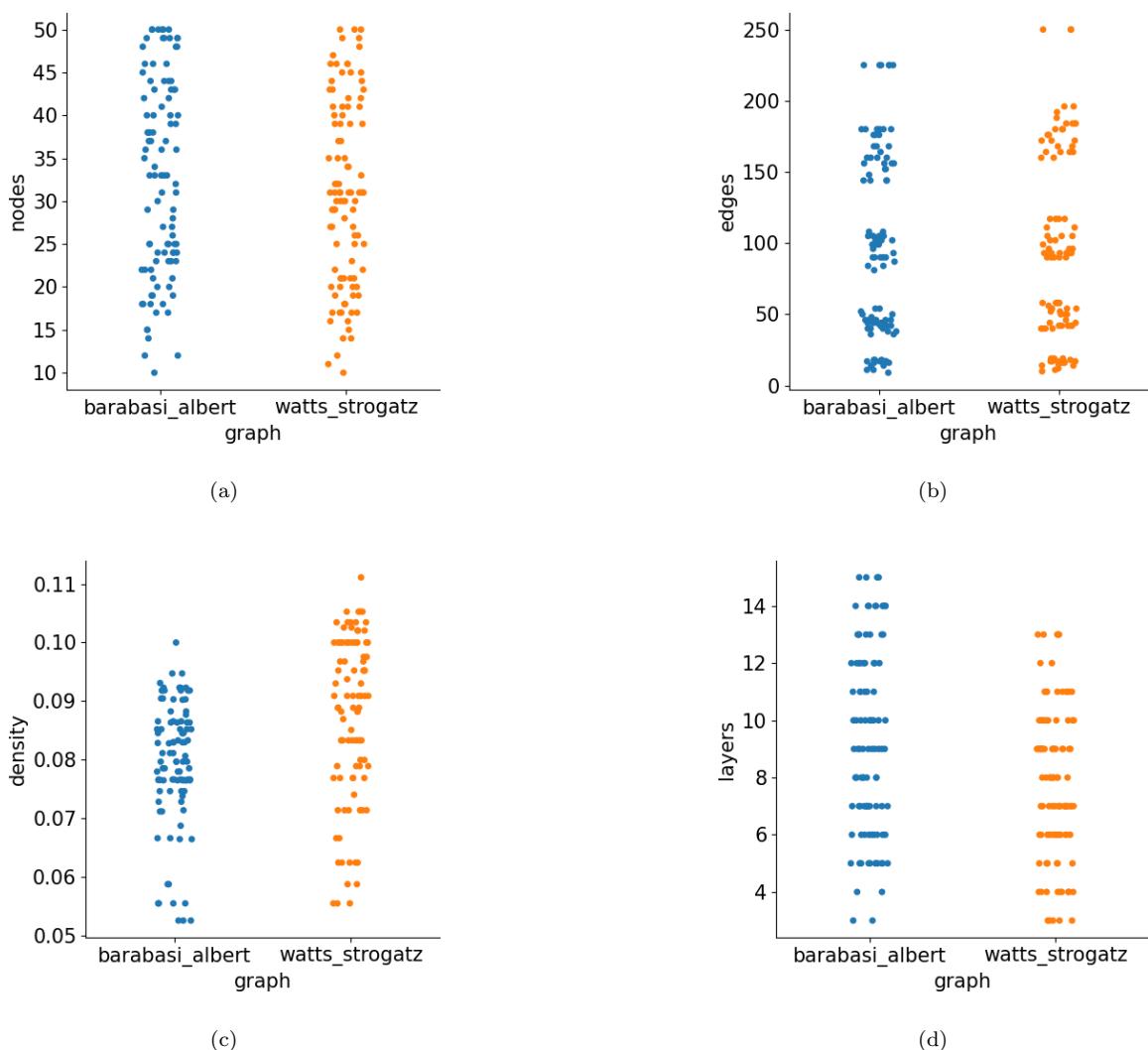


Figure C.1: Comparison of basic graph properties and number of layers in Watts–Strogatz and Barabási–Albert based RNN_Tanh models

C Plots

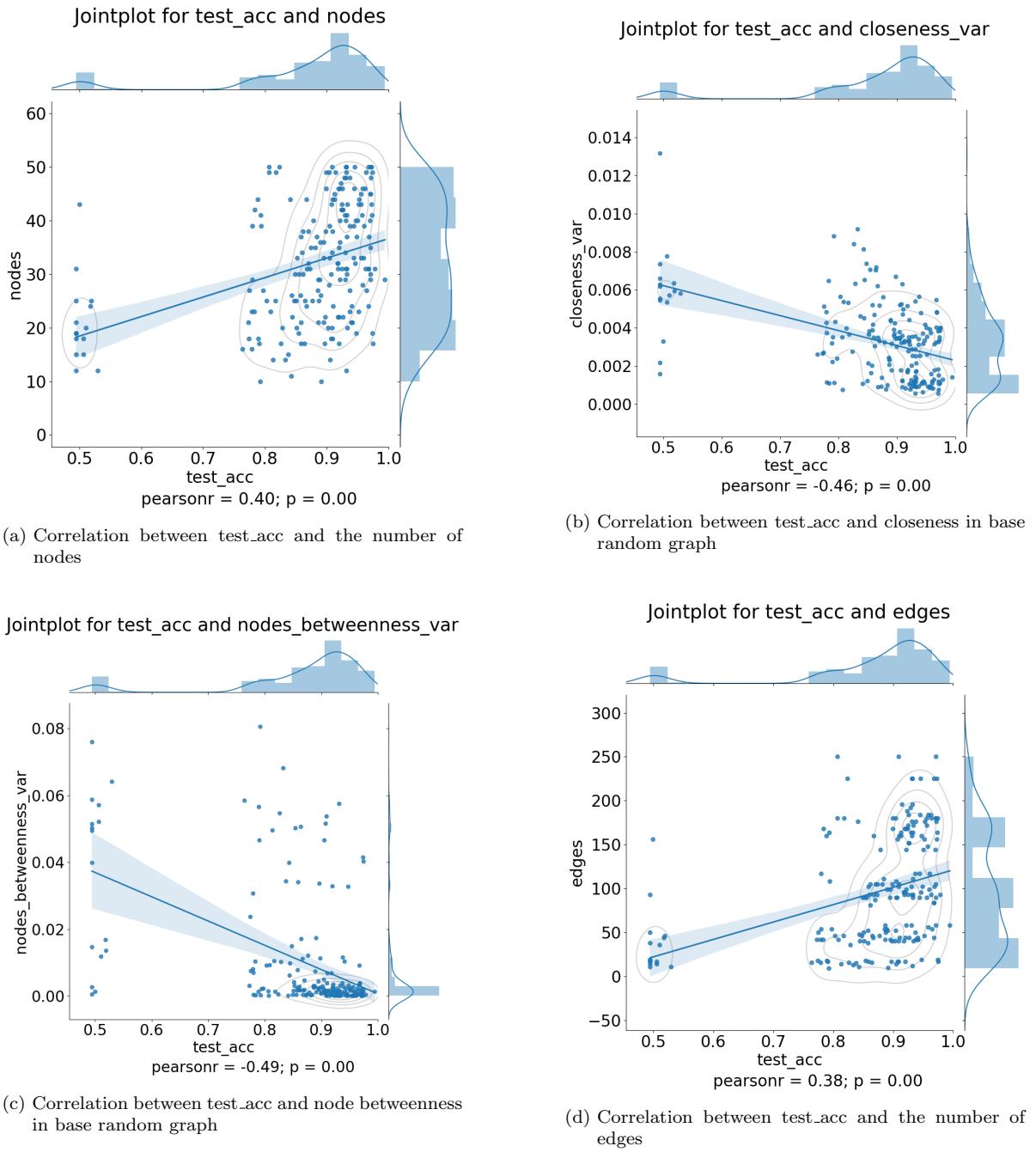


Figure C.2: Correlation between test accuracy of RNN_Tanh and its different graph and recurrent network properties

C.2 Randomly structured RNN_ReLU

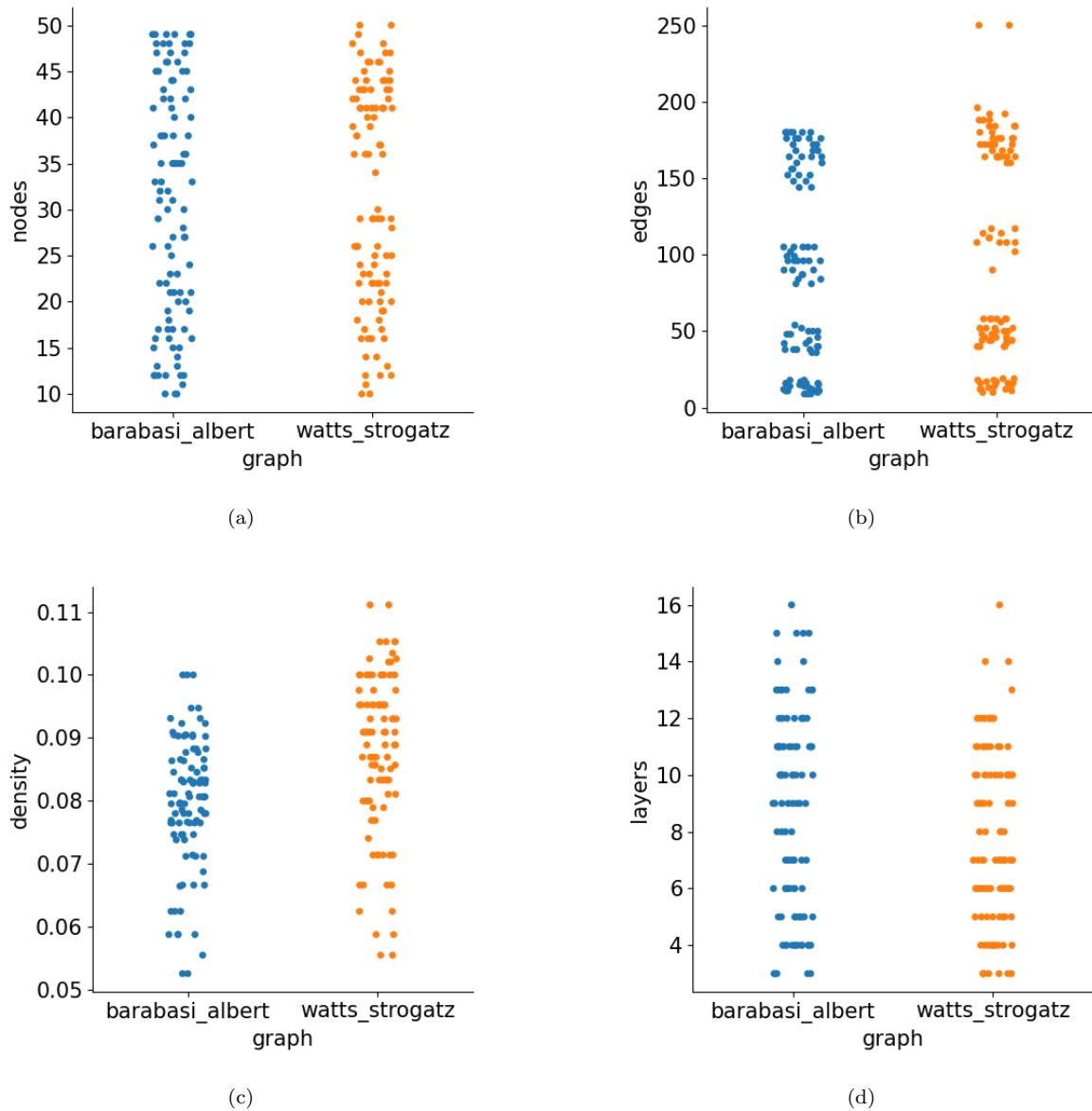
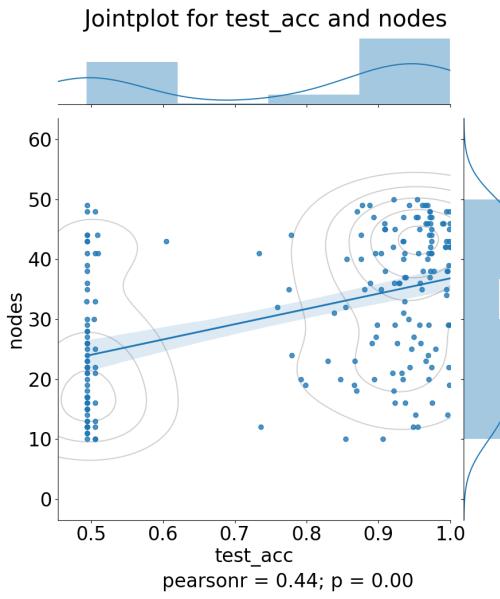
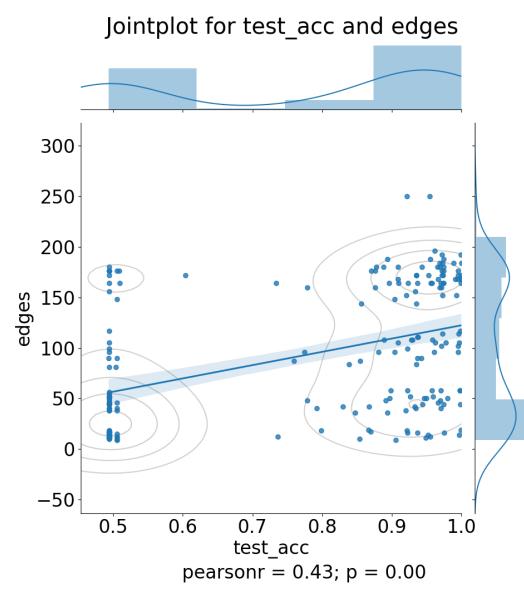


Figure C.3: Comparison of basic graph properties and number of layers in Watts–Strogatz and Barabási–Albert based RNN_ReLU models

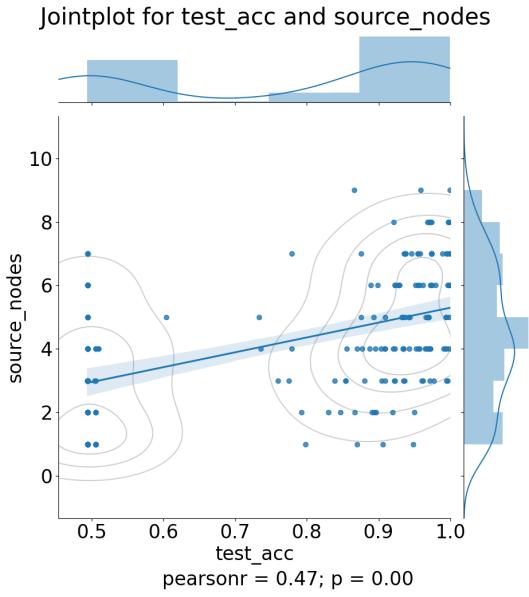
C Plots



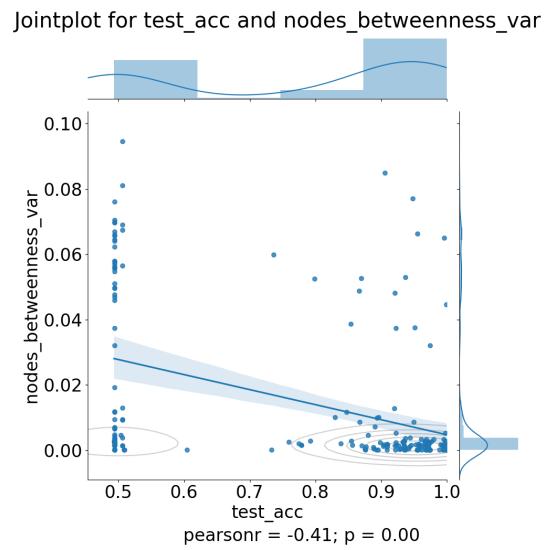
(a) Correlation between test_acc and the number of nodes



(b) Correlation between test_acc and the number of edges



(c) Correlation between test_acc and the number of source nodes



(d) Correlation between test_acc and node betweenness in base random graph

Figure C.4: Correlation between test accuracy of RNN_ReLU and its different graph and recurrent network properties

C.3 Randomly structured LSTM

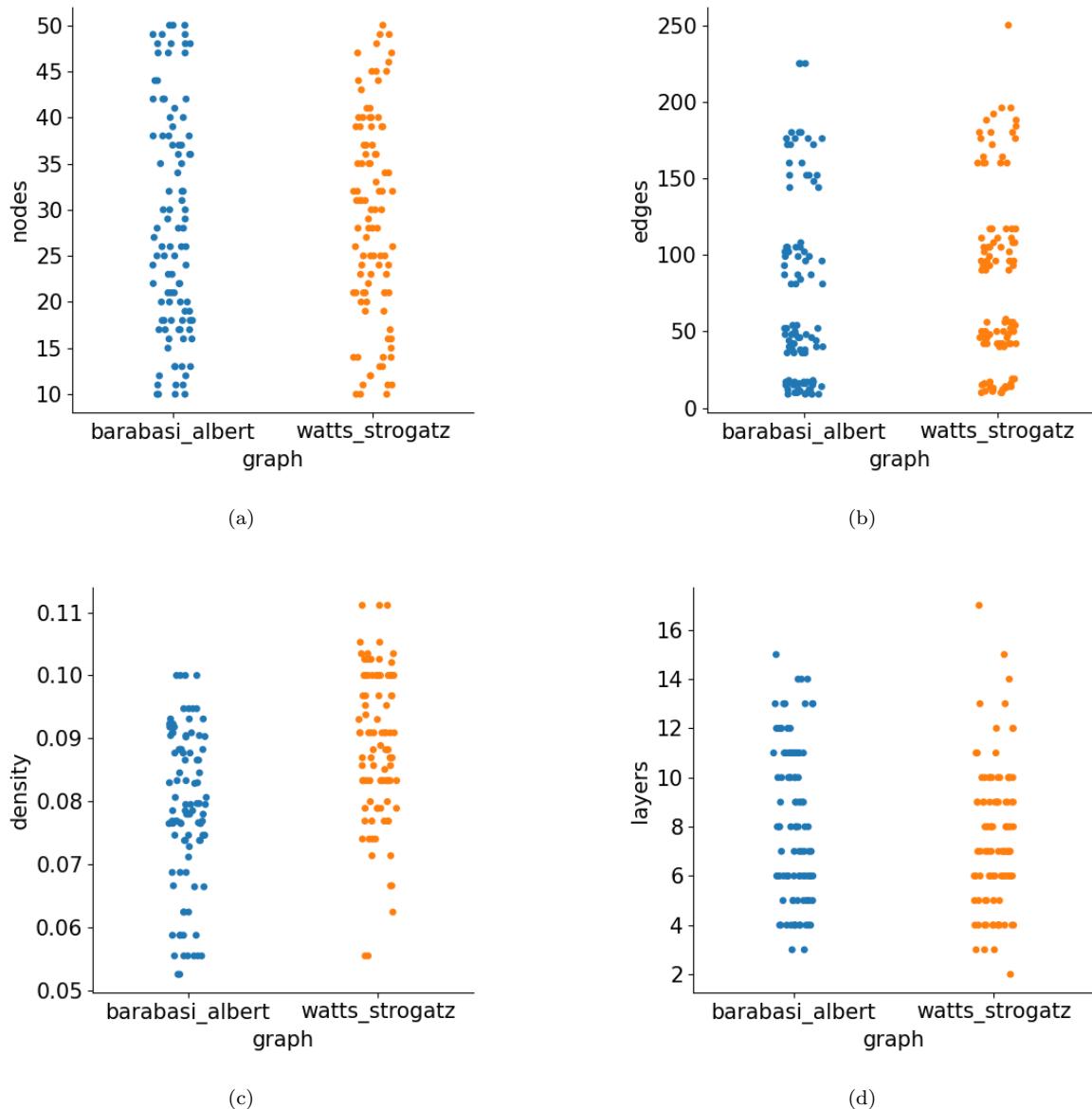


Figure C.5: Comparison of basic graph properties and number of layers in Watts–Strogatz and Barabási–Albert based LSTM models

C Plots

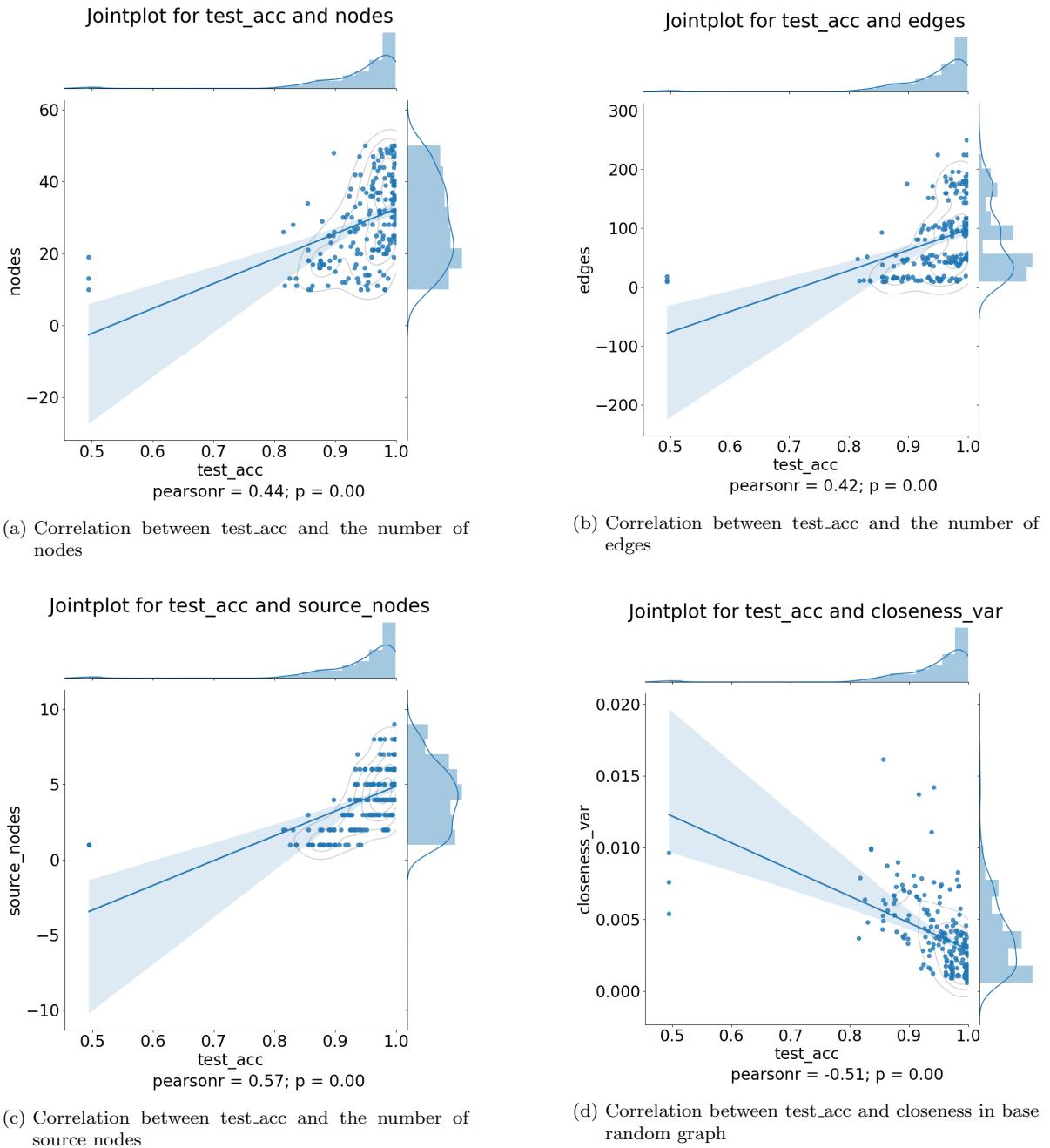
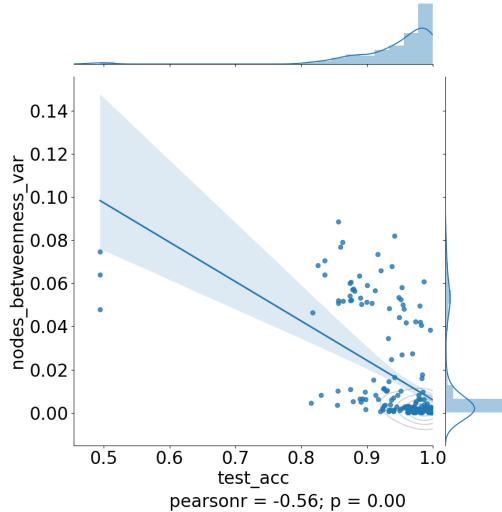


Figure C.6: Correlation between test accuracy of LSTM and its different graph and recurrent network properties - 1

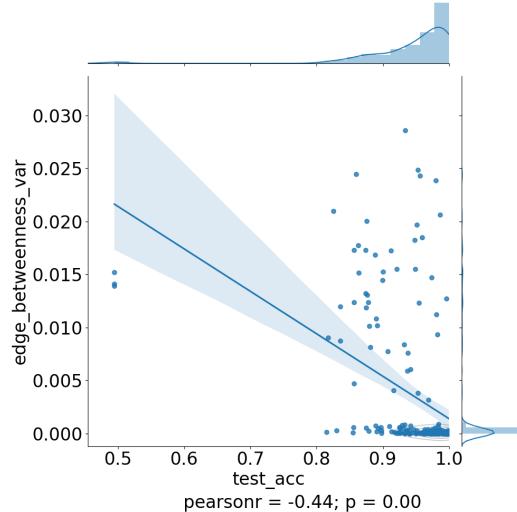
C Plots

Jointplot for test_acc and nodes_betweenness_var



(a) Correlation between test_acc and node betweenness in base random graph

Jointplot for test_acc and edge_betweenness_var



(b) Correlation between test_acc and edge betweenness in base random graph

Figure C.7: Correlation between test accuracy of LSTM and its different graph and recurrent network properties - 2

C.4 Randomly structured GRU

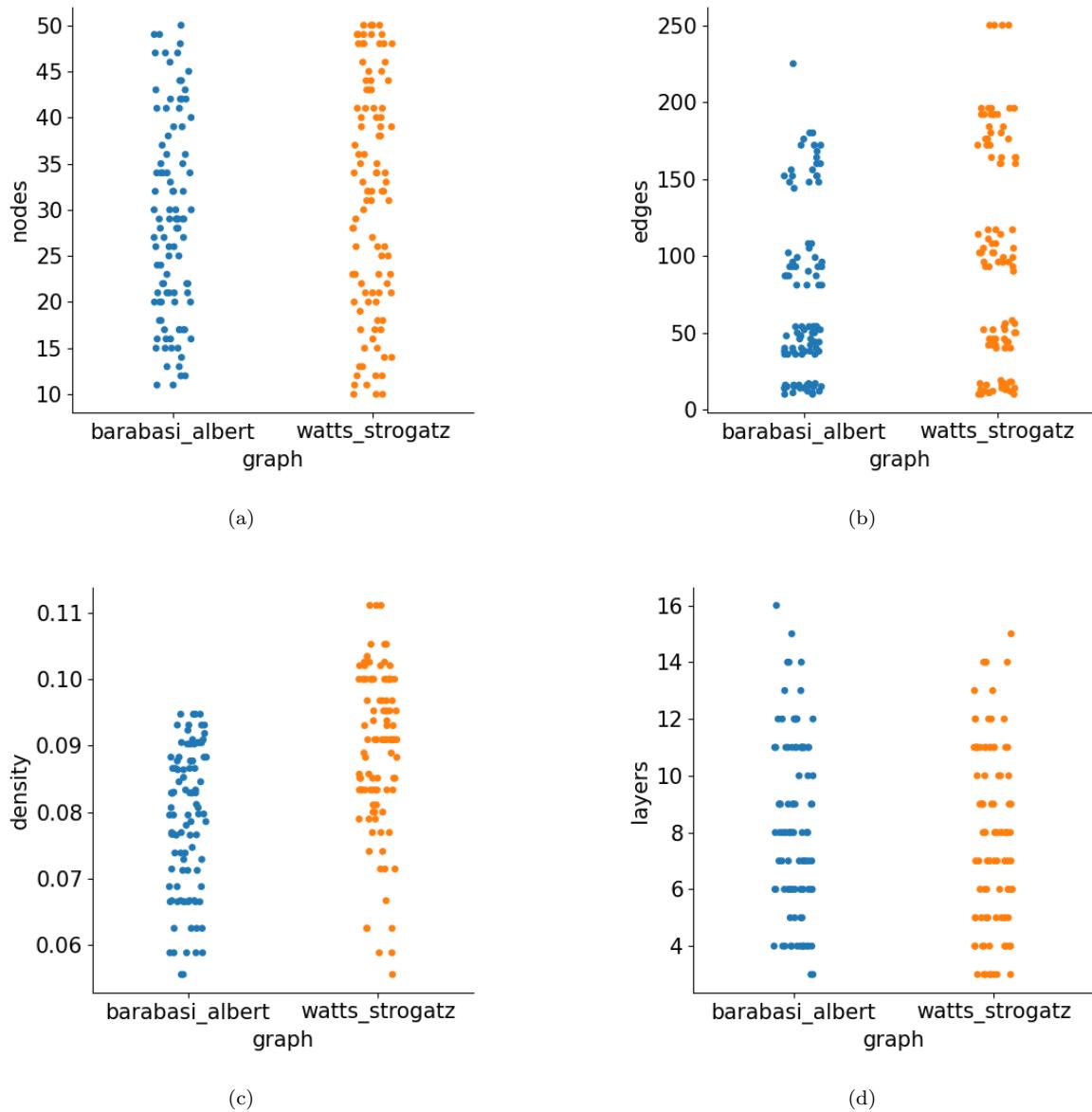
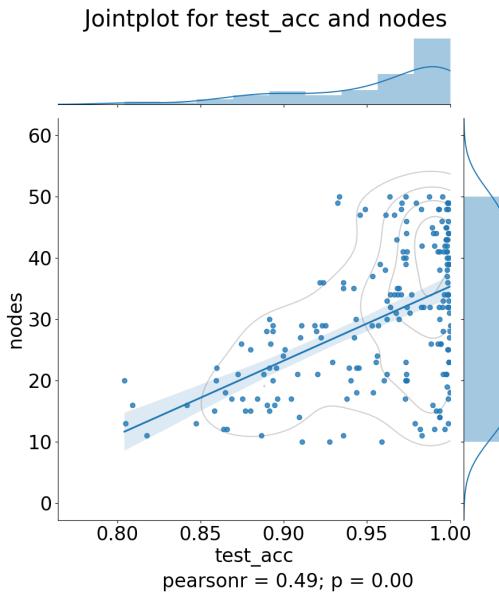
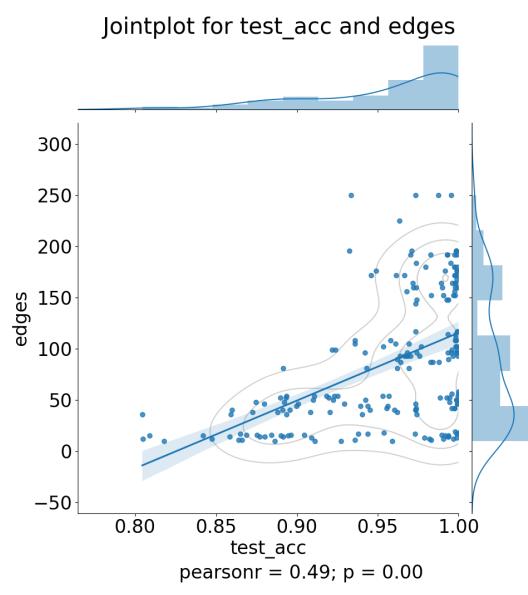


Figure C.8: Comparison of basic graph properties and number of layers in Watts–Strogatz and Barabási–Albert based GRU models

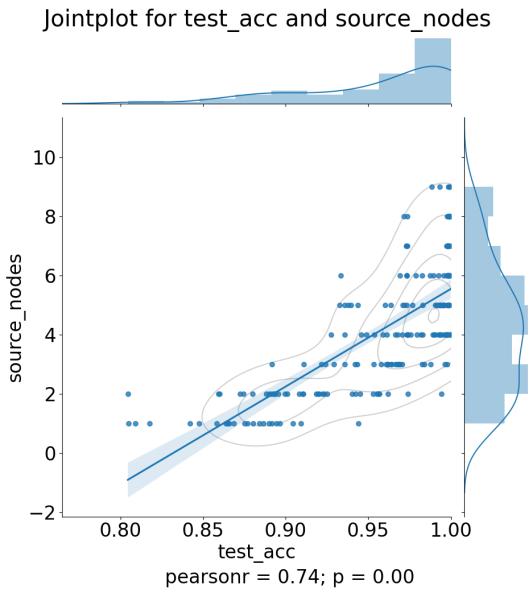
C Plots



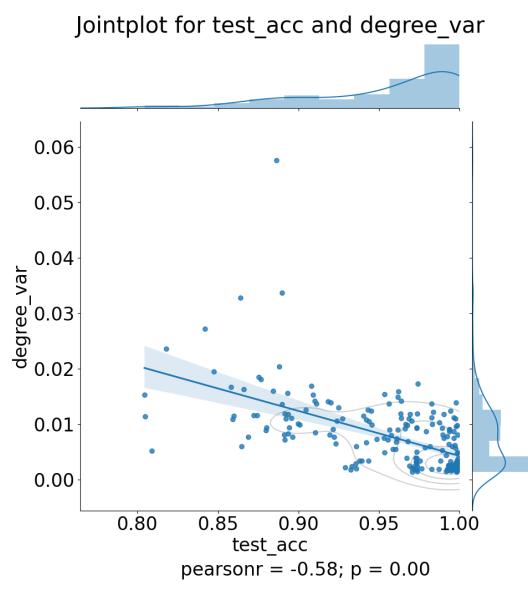
(a) Correlation between test_acc and the number of nodes



(b) Correlation between test_acc and the number of edges



(c) Correlation between test_acc and the number of source nodes



(d) Correlation between test_acc and degree of the base random graph

Figure C.9: Correlation between test accuracy of GRU and its different graph and recurrent network properties - 1

C Plots

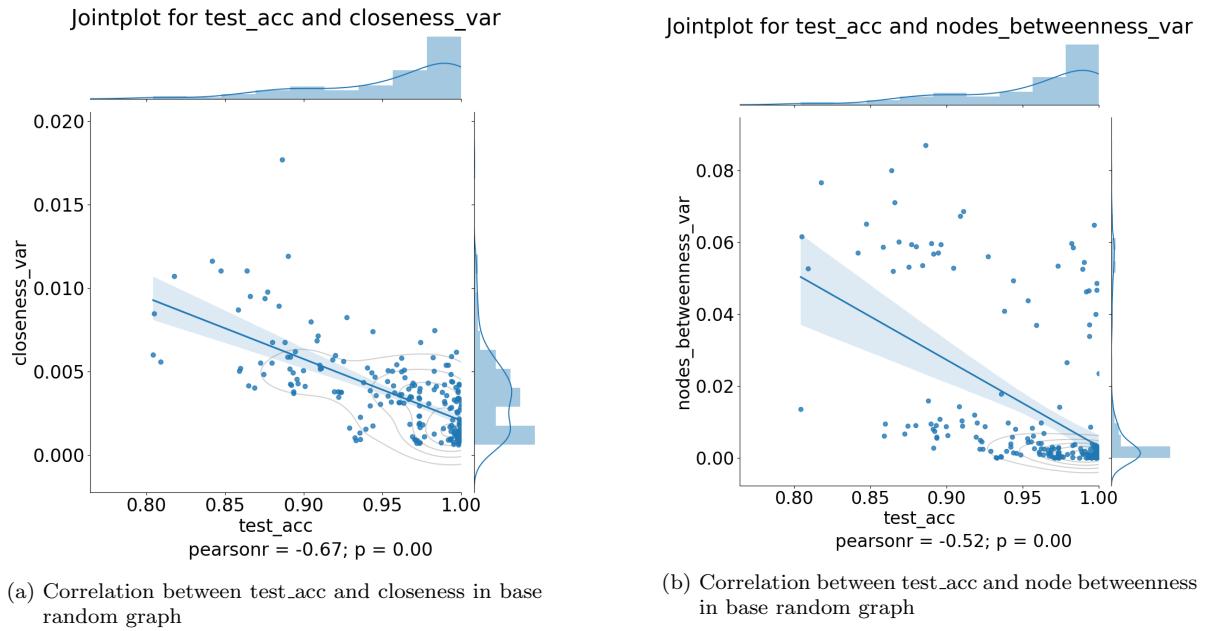


Figure C.10: Correlation between test accuracy of GRU and its different graph and recurrent network properties - 2

Bibliography

- [ANH13] J. Aarstad, H. Ness, and S. A. Haugland. “In what ways are small-world and scale-free networks interrelated?” In: *2013 IEEE International Conference on Industrial Technology (ICIT)*. 2013, pp. 1483–1487. DOI: [10.1109/ICIT.2013.6505891](https://doi.org/10.1109/ICIT.2013.6505891) (cit. on p. 34).
- [Alf+18] Simon Alford et al. “Pruned and Structurally Sparse Neural Networks”. In: *CoRR* abs/1810.00299 (2018). arXiv: [1810 . 00299](https://arxiv.org/abs/1810.00299). URL: <http://arxiv.org/abs/1810.00299> (cit. on p. 2).
- [BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. cite arxiv:1409.0473 Comment: Accepted at ICLR 2015 as oral presentation. 2014. URL: <http://arxiv.org/abs/1409.0473> (cit. on p. 1).
- [Bla+20] Davis Blalock et al. *What is the State of Neural Network Pruning?* 2020. arXiv: [2003 . 03033 \[cs.LG\]](https://arxiv.org/abs/2003.03033) (cit. on p. 29).
- [Car20] Stephan C. Carlson. “Graph theory”. In: (Nov. 2020). URL: <https://www.britannica.com/topic/graph-theory> (cit. on p. 30).
- [Che+13] Ciprian Chelba et al. *One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling*. Tech. rep. Google, 2013. URL: <http://arxiv.org/abs/1312.3005> (cit. on p. 8).
- [Cho+14] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014) (cit. on p. 27).
- [Dey+19] S. Dey et al. “Pre-Defined Sparse Neural Networks With Hardware Acceleration”. In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9.2 (2019), pp. 332–345. DOI: [10 . 1109 / JETCAS . 2019 . 2910864](https://doi.org/10.1109/JETCAS.2019.2910864) (cit. on p. 2).

Bibliography

- [ES02] Douglas Eck and Jürgen Schmidhuber. “Learning the Long-Term Structure of the Blues”. In: *Artificial Neural Networks — ICANN 2002*. Ed. by José R. Dorronsoro. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 284–289. ISBN: 978-3-540-46084-8 (cit. on p. 26).
- [Fuk80] Kunihiko Fukushima. “Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position”. In: *Biological Cybernetics* 36 (1980), pp. 193–202 (cit. on p. 19).
- [GG16] Yuan Gao and Dorota Glowacka. “Deep gate recurrent neural network”. In: *Asian conference on machine learning*. 2016, pp. 350–365 (cit. on p. 27).
- [GO94] C Lee Giles and Christian W Omlin. “Pruning recurrent neural networks for improved generalization performance”. In: *IEEE transactions on neural networks* 5.5 (1994), pp. 848–851 (cit. on p. 6).
- [GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016 (cit. on p. 15).
- [Gup+18] Anvita Gupta et al. “Generative Recurrent Networks for De Novo Drug Design”. In: *Molecular Informatics* 37.1-2 (2018), p. 1700111. DOI: <https://doi.org/10.1002/minf.201700111>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.201700111>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201700111> (cit. on p. 26).
- [HQ12] Hong-Gui Han and Jun-Fei Qiao. “Prediction of activated sludge bulking based on a self-organizing RBF neural network”. In: *Journal of Process Control* 22.6 (2012), pp. 1103–1112 (cit. on p. 7).
- [HZQ17] Hong-Gui Han, Shuo Zhang, and Jun-Fei Qiao. “An Adaptive Growing and Pruning Algorithm for Designing Recurrent Neural Network”. In: *Neurocomput.* 242.C (June 2017), pp. 51–62. ISSN: 0925-2312. DOI: [10.1016/j.neucom.2017.02.038](https://doi.org/10.1016/j.neucom.2017.02.038). URL: <https://doi.org/10.1016/j.neucom.2017.02.038> (cit. on p. 6).
- [He+16] K. He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90) (cit. on p. 2).

Bibliography

- [HS13] Michiel Hermans and Benjamin Schrauwen. “Training and Analyzing Deep Recurrent Neural Networks”. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. NIPS’13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 190–198 (cit. on p. 21).
- [HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780 (cit. on pp. 25, 36).
- [HYJ14] Han Honggui, Li Ying, and Qiao Junfei. “A fuzzy neural network approach for online fault detection in waste water treatment process”. In: *Computers & Electrical Engineering* 40.7 (2014), pp. 2216–2226 (cit. on p. 6).
- [Jor86] Michael I. Jordan. *Serial Order: A Parallel, Distributed Processing Approach*. Tech. rep. 8604. Institute for Cognitive Science, University of California, San Diego, 1986 (cit. on p. 20).
- [KO11] Bekir Karlik and A Vehbi Olgac. “Performance analysis of various activation functions in generalized MLP architectures of neural networks”. In: *International Journal of Artificial Intelligence and Expert Systems* 1.4 (2011), pp. 111–122 (cit. on p. 12).
- [LHL20] Xuan-Hien Le, Hung Viet Ho, and Giha Lee. “Application of Gated Recurrent Unit (GRU) Network for Forecasting River Water Levels Affected by Tides”. In: *APAC 2019*. Ed. by Nguyen Trung Viet, Dou Xiping, and Tran Thanh Tung. Singapore: Springer Singapore, 2020, pp. 673–680. ISBN: 978-981-15-0291-0 (cit. on p. 28).
- [Lec+98] Y. Lecun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: [10.1109/5.726791](https://doi.org/10.1109/5.726791) (cit. on pp. 1, 7, 8).
- [LAT18] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip HS Torr. “Snip: Single-shot network pruning based on connection sensitivity”. In: *arXiv preprint arXiv:1810.02340* (2018) (cit. on p. 8).
- [Li+18] Guiying Li et al. “Optimization based Layer-wise Magnitude-based Pruning for DNN Compression.” In: *IJCAI*. Vol. 1. 2018, p. 2 (cit. on p. 29).
- [Lia+16] Z. Liang et al. “CNN-based image analysis for malaria diagnosis”. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2016, pp. 493–496. DOI: [10.1109/BIBM.2016.7822567](https://doi.org/10.1109/BIBM.2016.7822567) (cit. on p. 19).

Bibliography

- [Liu+15] Baoyuan Liu et al. “Sparse Convolutional Neural Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015 (cit. on p. 2).
- [Mao+17] Huizi Mao et al. “Exploring the regularity of sparse structure in convolutional neural networks”. In: *arXiv preprint arXiv:1705.08922* (2017) (cit. on p. 2).
- [MP43] Warren Mcculloch and Walter Pitts. “A Logical Calculus of Ideas Immanent in Nervous Activity”. In: *Bulletin of Mathematical Biophysics* 5 (1943), pp. 127–147 (cit. on p. 9).
- [ML15] Meiyin Wu and Li Chen. “Image recognition based on deep learning”. In: *2015 Chinese Automation Congress (CAC)*. 2015, pp. 542–546. DOI: [10.1109/CAC.2015.7382560](https://doi.org/10.1109/CAC.2015.7382560) (cit. on p. 19).
- [Men18] Ahmed Menshawy. *Deep Learning By Example*. Packt Publishing, 2018 (cit. on p. 12).
- [Mer+16] Stephen Merity et al. “Pointer Sentinel Mixture Models”. In: *CoRR* abs/1609.07843 (2016). arXiv: [1609.07843](https://arxiv.org/abs/1609.07843). URL: <http://arxiv.org/abs/1609.07843> (cit. on p. 8).
- [NUD17] Sharan Narang, Eric Undersander, and Gregory Diamos. “Block-sparse recurrent neural networks”. In: *arXiv preprint arXiv:1711.02782* (2017) (cit. on p. 7).
- [Nar+17] Sharan Narang et al. “Exploring Sparsity in Recurrent Neural Networks”. In: *CoRR* abs/1704.05119 (2017). arXiv: [1704.05119](https://arxiv.org/abs/1704.05119). URL: <http://arxiv.org/abs/1704.05119> (cit. on p. 7).
- [Nea92] Radford M. Neal. “Connectionist Learning of Belief Networks”. In: *Artif. Intell.* 56.1 (July 1992), pp. 71–113. ISSN: 0004-3702. DOI: [10.1016/0004-3702\(92\)90065-6](https://doi.org/10.1016/0004-3702(92)90065-6). URL: [https://doi.org/10.1016/0004-3702\(92\)90065-6](https://doi.org/10.1016/0004-3702(92)90065-6) (cit. on p. 12).
- [Nwa+18] Chigozie Nwankpa et al. “Activation Functions: Comparison of trends in Practice and Research for Deep Learning”. In: *CoRR* abs/1811.03378 (2018). arXiv: [1811.03378](https://arxiv.org/abs/1811.03378). URL: <http://arxiv.org/abs/1811.03378> (cit. on p. 12).

Bibliography

- [PBW17] Tulasi K. Paradarami, Nathaniel D. Bastian, and Jennifer L. Wightman. “A hybrid recommender system using artificial neural networks”. In: *Expert Systems with Applications* 83 (2017), pp. 300–313. ISSN: 0957-4174. DOI: <https://doi.org/10.1016/j.eswa.2017.04.046>. URL: <http://www.sciencedirect.com/science/article/pii/S0957417417302968> (cit. on p. 2).
- [Por12] M. A. Porter. “Small-world network”. In: *Scholarpedia* 7.2 (2012). revision #187873, p. 1739. DOI: [10.4249/scholarpedia.1739](https://doi.org/10.4249/scholarpedia.1739) (cit. on p. 33).
- [Ras15] Sebastian Raschka. *Single-Layer Neural Networks and Gradient Descent*. 2015. URL: https://sebastianraschka.com/Articles/2015_singelayer_neurons.html (cit. on p. 13).
- [Rav+18] M. Ravanelli et al. “Light Gated Recurrent Units for Speech Recognition”. In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 2.2 (2018), pp. 92–102. DOI: [10.1109/TETCI.2017.2762739](https://doi.org/10.1109/TETCI.2017.2762739) (cit. on p. 28).
- [Reb76] Arthur S Reber. “Implicit learning of synthetic languages: The role of instructional set.” In: *Journal of Experimental Psychology: Human Learning and Memory* 2.1 (1976), p. 88 (cit. on p. 36).
- [Res17] Ronny Restrepo. *Derivative of the Sigmoid function - a worked example*. 2017. URL: http://ronny.rest/blog/post_2017_08_10_sigmoid/ (cit. on p. 11).
- [Ros58] F. Rosenblatt. “The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain”. In: *Psychological Review* (1958), pp. 65–386 (cit. on p. 9).
- [RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning Representations by Back-propagating Errors”. In: *Nature* 323.6088 (1986), pp. 533–536. DOI: [10.1038/323533a0](https://doi.org/10.1038/323533a0). URL: <http://www.nature.com/articles/323533a0> (cit. on p. 16).
- [Sch13] Falk Schreiber. “Characteristic Path Length”. In: *Encyclopedia of Systems Biology*. Ed. by Werner Dubitzky et al. New York, NY: Springer New York, 2013, pp. 395–395. ISBN: 978-1-4419-9863-7. DOI: [10.1007/978-1-4419-9863-7_1460](https://doi.org/10.1007/978-1-4419-9863-7_1460). URL: https://doi.org/10.1007/978-1-4419-9863-7_1460 (cit. on p. 33).

Bibliography

- [STS18] Sima Siami Namini, Neda Tavakoli, and Akbar Siami Namin. “A Comparison of ARIMA and LSTM in Forecasting Time Series”. In: Dec. 2018, pp. 1394–1401. DOI: [10.1109/ICMLA.2018.00227](https://doi.org/10.1109/ICMLA.2018.00227) (cit. on p. 26).
- [SG19] Julian Stier and Michael Granitzer. “Structural Analysis of Sparse Neural Networks”. In: *CoRR* abs/1910.07225 (2019). arXiv: [1910.07225](https://arxiv.org/abs/1910.07225). URL: <http://arxiv.org/abs/1910.07225> (cit. on pp. 4, 7, 47).
- [SLC17] Qinliang Su, Xuejun Liao, and Lawrence Carin. “A Probabilistic Framework for Nonlinearities in Stochastic Neural Networks”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4489–4498. ISBN: 9781510860964 (cit. on p. 10).
- [SWT16] Yi Sun, Xiaogang Wang, and Xiaoou Tang. “Sparsifying neural network connections for face recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4856–4864 (cit. on p. 2).
- [Sut13] Ilya Sutskever. “Training Recurrent Neural Networks”. PhD thesis. CAN, 2013. ISBN: 9780499220660 (cit. on p. 20).
- [TKT18] Boukaye Boubacar Traore, Bernard Kamsu-Foguem, and Fana Tangara. “Deep convolution neural network for image recognition”. In: *Ecological Informatics* 48 (2018), pp. 257–268. ISSN: 1574-9541. DOI: <https://doi.org/10.1016/j.ecoinf.2018.10.002>. URL: <http://www.sciencedirect.com/science/article/pii/S1574954118302140> (cit. on p. 1).
- [VMN20] Greg Van Houdt, Carlos Mosquera, and Gonzalo Nápoles. “A Review on the Long Short-Term Memory Model”. In: *Artificial Intelligence Review* 53 (Dec. 2020). DOI: [10.1007/s10462-020-09838-1](https://doi.org/10.1007/s10462-020-09838-1) (cit. on p. 25).
- [WZG20] Chaoqi Wang, Guodong Zhang, and Roger Grosse. *Picking Winning Tickets Before Training by Preserving Gradient Flow*. 2020. arXiv: [2002.07376 \[cs.LG\]](https://arxiv.org/abs/2002.07376) (cit. on p. 8).
- [WS98] Duncan J Watts and Steven H Strogatz. “Collective dynamics of ‘small-world’ networks”. In: *nature* 393.6684 (1998), pp. 440–442 (cit. on p. 34).
- [Wei] Eric W Weisstein. *Lattice Graph*. URL: <https://mathworld.wolfram.com/LatticeGraph.html> (cit. on p. 34).

Bibliography

- [Wen+20] Liangjian Wen et al. “Structured pruning of recurrent neural networks through neuron selection”. In: *Neural Networks* 123 (2020), pp. 134–141 (cit. on p. 8).
- [Wer90] P. Werbos. “Backpropagation Through Time: What It Does and How to Do It”. In: 1990 (cit. on p. 21).
- [YYZ20] S. Yang, X. Yu, and Y. Zhou. “LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example”. In: *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*. 2020, pp. 98–101. doi: [10.1109/IWECAI50956.2020.00027](https://doi.org/10.1109/IWECAI50956.2020.00027) (cit. on p. 28).
- [YMJ12] Jun Yin, Yan Meng, and Yaochu Jin. “A developmental approach to structural self-organization in reservoir computing”. In: *IEEE transactions on autonomous mental development* 4.4 (2012), pp. 273–289 (cit. on p. 7).
- [ZS19] Matthew Shunshi Zhang and Bradly Stadie. “One-Shot Pruning of Recurrent Neural Networks by Jacobian Spectrum Evaluation”. In: *CoRR* abs/1912.00120 (2019). arXiv: [1912 . 00120](https://arxiv.org/abs/1912.00120). URL: <http://arxiv.org/abs/1912.00120> (cit. on p. 8).

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich diese Masterarbeit selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe und alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, als solche gekennzeichnet sind, sowie, dass ich die Masterarbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt habe.

Passau, July 1, 2021

Harshil Jagadishbhai Darji