GRAPH EXPANSION IN PRUNED RECURRENT NEURAL NETWORK LAYERS PRESERVES PERFORMANCE

Suryam Arnav Kalra *
Computer Science and Engineering, IIT Kharagpur

Arindam Biswas
Research Scientist, Polynom, France

Pabitra Mitra
Computer Science and Engineering, IIT Kharagpur

Biswajit Basu
Engineering School, Trinity College Dublin

ABSTRACT

Expansion property of a graph refers to its strong connectivity as well as sparseness. We prune recurrent networks such as RNNs and LSTMs, maintaining a large spectral gap of the underlying graphs and ensuring their layer-wise expansion properties. We also study the time unfolded recurrent network graphs in terms of the properties of their bipartite layers. Experimental results for the benchmark sequence datasets (along with noise) show that expander graph properties are key to preserving classification accuracy of RNN and LSTM.

1 INTRODUCTION

Neural networks can often be pruned to very high sparsity while maintaining the performance. This phenomenon has been stated as the lottery ticket hypothesis (Frankle & Carbin, 2018). It has been observed that the winning lottery tickets follow certain desirable graph theoretic properties. The relation between the lottery ticket hypothesis and expander and Ramanujan graph properties for fully connected and convolutional neural networks has been previously explored in (Pal et al., 2022).

Expander properties of feed-forward networks in general have been well studied in the literature (Prabhu et al., 2018; Hoang et al., 2023; Stewart et al., 2023; Esguerra et al., 2023). However, there is no work which studies the performance of recurrent networks like RNNs and LSTMs with respect to their expansion properties. In this paper we study the expansion properties of recurrent neural networks (RNN) and LSTM, and observe that performance of such networks is strongly correlated with the spectral bounds characterizing the properties. We adopt a method for time unrolling the recurrent structures to obtain bipartite graphs on which spectral bounds are computed.

2 RAMANUJAN BOUNDS AND NETWORK STRUCTURES OF RNN AND LSTM

Let $\Gamma = (V, E)$ be a d-regular $(d \geq 3)$ bipartite graph. Let the eigenvalues of its adjacency matrix be $\lambda_n \leq \lambda_{n-1} \leq \ldots \leq \lambda_2 \leq \lambda_1$. Then Γ is said to be Ramanujan iff $|\lambda_i| \leq 2\sqrt{d-1}$, for $i = 2, \ldots, (n-1)$ (Lubotzky et al., 1988). The quantity d_{avg} is the average degree of all vertices. Using this, we consider the following expressions $\Delta_R = \frac{2\sqrt{d_{avg}-1}-\lambda_2}{\lambda_2}$ and $\Delta_S = \frac{2\sqrt{\lambda_1-1}-\lambda_2}{\lambda_2}$. The pruning process is depicted in 1. For details see Appendix.

Recurrent networks and LSTMs are cyclic structures which can be made acyclic by folding over time. In the unrolled network the weights are copied over the time steps, and only the hidden state and the input values change. Given a RNN or a LSTM, we consider the complete bipartite network through which the input data passes during inference. The graph structures are shown in Figure 2.

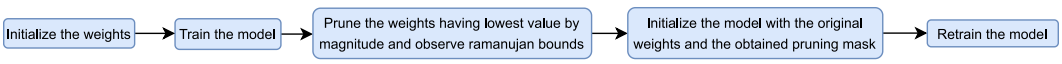
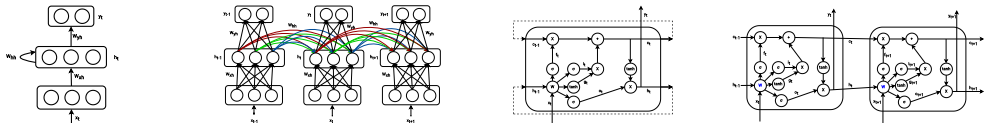


Figure 1: Pruning Process

*Corresponding Author



(a) Classic RNN (b) Unfolded RNN (c) Classic LSTM (d) Unfolded LSTM

Figure 2: Bipartite representation of RNN and LSTM by unfolding over the time steps

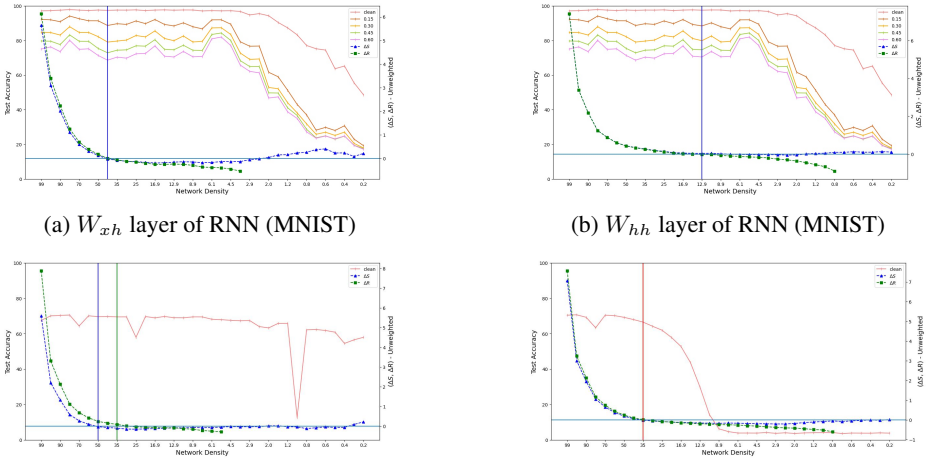
3 Data Sets and Experimental Results

We provide results for RNN trained on the sequential MNIST dataset ($k = 28$) (Le et al., 2015) and LSTM trained on the Speech Commands dataset ($k = 400$) (Warden, 2018) where k is the sequence length. Accuracy on the test split of the dataset (20%) is used as the performance metric. We also add Gaussian noise with 0 mean and σ variance to p fraction of the pixels of MNIST. We considered $p = 0.20$ and $\sigma = 0.15, 0.30, 0.45, 0.60$, respectively to study the effect of varying degree of noise.

Table 1: Hyperparameters for the experiments

Learning Rate	Training Epochs	Pruning Epochs	Batch Size	Optimizer	Initialization	Loss Function
0.001	20	20	128	Adam	Kaiming Uniform	Cross Entropy

The goal of our experiments is to study the effect of preserving expander graph properties on the performance of sparse RNN and LSTM. One shot pruning is used to sparsify the network. The weights between input to hidden layers, feedback layers, and hidden to output layers are represented as W_{xh}, W_{hh}, and W_{hy}. We only prune the W_{xh} and W_{hh} layers, while the W_{hy} layer is unchanged. For MNIST, the W_{xh} and W_{hh} weights lose the Ramanujan property at a network density of 50.0% and 12.0% respectively. For the speech command dataset the zero crossing is at network density of 50% and 35% for W_{xh} and 35% for W_{hh}. Figure 3 shows that the degradation in performance with loss of Ramanujan property is even more prominent as noise increases. This reinstates the fact that Ramanujan property is crucial for noise robustness of the networks. We observe in most of our experiments with RNN that the W_{xh} layer lose the expander graph property before the W_{hh} layer. This points to the fact that W_{xh} edges play more significant role as compared to the W_{hh} ones.



(a) W_{xh} layer of RNN (MNIST) (b) W_{hh} layer of RNN (MNIST)

(c) W_{xh} layer of LSTM (Speech) (d) W_{hh} layer of LSTM (Speech)

Figure 3: Variation in test accuracy and spectral gap (Δ_S, Δ_R) of unweighted graph representation.

4 Conclusion

Maintaining the high expansion property is a sufficient condition to obtain a high accuracy. When the Ramanujan graph bounds are satisfied, the accuracy doesn't drop. Once the expansion falls below the Ramanujan graph bound, accuracy is maintained for a short while, before it eventually collapses. Eventually the network becomes weakly connected and the accuracy falls.

URM DECLARATION

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Noga Alon and Vitali D Milman. $\lambda 1$, isoperimetric inequalities for graphs, and superconcentrators. *Journal of Combinatorial Theory, Series B*, 38(1):73–88, 1985.

Arindam Biswas. On a cheeger type inequality in cayley graphs of finite groups. *European Journal of Combinatorics*, 81:298–308, October 2019. doi: 10.1016/j.ejc.2019.06.009. URL https://doi.org/10.1016/j.ejc.2019.06.009.

Arindam Biswas and Jyoti Prakash Saha. A Cheeger type inequality in finite Cayley sum graphs. *Algebraic Combinatorics*, 4(3):517–531, 2021. doi: 10.5802/alco.166. URL https://alco.centre-mersenne.org/articles/10.5802/alco.166/.

Arindam Biswas and Jyoti Prakash Saha. Expansion in Cayley graphs, Cayley sum graphs and their twists. *arXiv e-prints*, art. arXiv:2103.05935, March 2021. doi: 10.48550/arXiv.2103.05935.

Arindam Biswas and Jyoti Prakash Saha. Spectra of twists of cayley and cayley sum graphs. *Advances in Applied Mathematics*, 132:102272, January 2022. doi: 10.1016/j.aam.2021.102272. URL https://doi.org/10.1016/j.aam.2021.102272.

Arindam Biswas and Jyoti Prakash Saha. A spectral bound for vertex-transitive graphs and their spanning subgraphs. *Algebraic Combinatorics*, 6(3):689–706, 2023. doi: 10.5802/alco.278. URL https://alco.centre-mersenne.org/articles/10.5802/alco.278/.

Emmanuel Breuillard, Ben Green, Robert Guralnick, and Terence Tao. Expansion in finite simple groups of lie type. *Journal of the European Mathematical Society*, 17(6):1367–1434, 2015. doi: 10.4171/jems/533. URL https://doi.org/10.4171/jems/533.

Fan Chung. A generalized alon-boppana bound and weak ramanujan graphs. *The Electronic Journal of Combinatorics*, 23(3), July 2016. doi: 10.37236/5933. URL https://doi.org/10.37236/5933.

Jozef Dodziuk. Difference equations, isoperimetric inequality and transience of certain random walks. *Transactions of the American Mathematical Society*, 284(2):787–794, 1984.

Kiara Esguerra, Muneeb Nasir, Tong Boon Tang, Afidalina Tumian, and Eric Tatt Wei Ho. Sparsity-aware orthogonal initialization of deep neural networks. *IEEE Access*, 2023.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Duc Hoang, Shiwei Liu, Radu Marculescu, and Zhangyang Wang. Revisiting pruning at initialization through the lens of ramanujan graph. In *International Conference on Learning Representations*, 2023.

Shlomo Hoory. A lower bound on the spectral radius of the universal cover of a graph. *Journal of Combinatorial Theory, Series B*, 93(1):33–43, January 2005. doi: 10.1016/j.jctb.2004.06.001. URL https://doi.org/10.1016/j.jctb.2004.06.001.

Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

Quoc V Le, Navdeep Jaitly, and Geoffrey E Hinton. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*, 2015.

A Lubotzky, R Phillips, and P Sarnak. Ramanujan graphs. *Combinatorica*, 8:261–277, 1988.

Bithika Pal, Arindam Biswas, Sudeshna Kolay, Pabitra Mitra, and Biswajit Basu. A study on the ramanujan graph property of winning lottery tickets. In *International Conference on Machine Learning, ICML*, volume 162, pp. 17186–17201, 2022.

Ameya Prabhu, Girish Varma, and Anoop Namboodiri. Deep expander networks: Efficient deep networks from graph theory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 20–35, 2018.

James Stewart, Umberto Michieli, and Mete Ozay. Data-free model pruning at initialization via expanders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4518–4523, 2023.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition, 2018.

A Appendix

A.1 Expander Graphs and Ramanujan Graphs

In this section, we discuss various properties of expanders which will be pertinent for the rest of the article. An expander graph is a sparse graph that has strong connectivity properties. The connectivity can be quantified in different ways giving rise to different notions of expanders such as vertex expanders, edge expanders and spectral expanders. These notions are interrelated. Recall that a graph $\Gamma = (V, E)$ is a tuple consisting of a vertex set V and an edge set E which is a subset of $V \times V$.

A.1.1 Combinatorial Expansion

Definition 1 (Expander and Cheeger constant) *A graph $\Gamma = (V, E)$ is said to be an ϵ-vertex expander if for every non-empty subset $X \subset V$ with $|X| \leq \frac{|V|}{2}$, we have $\frac{|\delta(X)|}{|X|} \geq \epsilon$, where $\delta(X)$ denotes the outer vertex boundary of X i.e., the set of vertices in Γ which are connected to a vertex in X but do not lie in X. The infimum as X runs over all subsets of V satisfying the conditions above is known as the vertex Cheeger constant and is denoted by $\mathfrak{h}(\Gamma)$.*

Edge expanders and the edge Cheeger constant $\mathbf{h}(\Gamma)$ are defined similarly, where in place of the vertex boundary, we consider the edge boundary i.e., the set of edges which have one vertex in X and the other outside of X. The vertex Cheeger constant $\mathfrak{h}(\Gamma)$ and the edge Cheeger constant $\mathbf{h}(\Gamma)$ are related by the following equivalence

$$\frac{\mathfrak{h}(\Gamma)}{D} \leqslant \mathbf{h}(\Gamma) \leqslant \mathfrak{h}(\Gamma),$$

where D denotes the maximum degree of the graph (the degree of each vertex is the number of edges going out from the vertex). The equivalence allows us to speak about vertex expansion and edge expansion interchangeably. Intuitively, given a graph with high vertex (or edge) Cheeger constant, it is more difficult to separate any subset of the vertices from the rest of the graph. This allows for free flow of information throughout the network which the graph modelises. In the literature, having a high Cheeger constant is also known as having high combinatorial expansion.

A.1.2 Spectral Expansion

The notion of spectral expansion is a bit different from combinatorial expansion. Given a finite undirected graph Γ the eigenvalues $\lambda_n \leqslant \cdots \leqslant \lambda_1$ of its adjacency matrix are all real and $\lambda_1 \leqslant D$ with equality iff the graph is D-regular. Recall that a graph is said to be d-regular if there are exactly d edges attached to a vertex. Thus, a d-regular bipartite graph is a graph which has the same number of vertices in each partition and every vertex of each partition has exactly d edges attached to it. A graph $\Gamma = (V, E)$ is said to be a spectral expander if the quantities $\{|\lambda_1| - |\lambda_2|, |\lambda_1| - |\lambda_k|\}$ are both bounded away from zero, where $k = n - 1$ if the graph is bipartite and $k = n$ otherwise.

A.2 DISCRETE CHEEGER–BUSER INEQUALITY

Ideally, to ensure free flow information within the network, our goal is to ensure that the graphs which modelise the networks have high combinatorial expansion. This is achieved via the discrete Cheeger–Buser inequality discovered independently by (Dodziuk, 1984) and by (Alon & Milman, 1985). The inequality states that

$$\frac{\mathbf{h}(\Gamma)^2}{2} \leqslant \alpha_2 \leqslant 2\mathbf{h}(\Gamma),$$

where α_2 denotes the second smallest eigenvalue of the normalised Laplacian matrix of Γ and is related to the eigenvalues of the adjacency matrix via

$$\frac{\lambda_i}{D} \leqslant 1 - \alpha_i \leqslant \frac{\lambda_i}{d} \; \forall i = 1, 2, \ldots, n.$$

See (Chung, 2016) for details. From the above, it is easy to check that a high $|\lambda_1| - |\lambda_2|$ ensures a high $\mathbf{h}(\Gamma)$ and vice-versa. Thus, the two notions of expansion are inter-connected and every spectral expander remains a combinatorial expander. They are actually equivalent for some classes of graphs, for instance bipartite graphs (as the adjacency spectrum is symmetric about the origin), variants of algebraic graphs (Breuillard et al., 2015; Biswas, 2019; Biswas & Saha, 2021; 2022; 2023; Biswas & Saha, 2021) etc.

A.2.1 RAMANUJAN GRAPH BOUNDS

A d-regular graph is said to be a Ramanujan graph if $\max\{|\lambda_2|, |\lambda_k|\} \leqslant 2\sqrt{d-1}$. In the case of bipartite graphs, $\lambda_k = \lambda_2$, hence the previous expression reduces to $|\lambda_2| \leqslant 2\sqrt{d-1}$. For fixed degree, with the sizes of the graphs growing larger and larger, these are the best possible expanders, as given by the Alon-Bopanna bound. We refer to Hoory–Linial–Wigderson (Hoory et al., 2006) for the details.

When the graphs modelising the network are irregular (and possibly weighted), to guarantee large expansion, we use two closely related quantities for d. The combinatorial quantity d_{avg} which is the average degree taking into account all vertices and the spectral quantity λ_1 which is the largest eigenvalue of the adjacency operator. The use of these quantities is justified by the work of Hoory (Hoory, 2005) and result in extremal families. Further, they have the added advantage of being easy to compute. Using them, we consider the following expressions Δ_R, Δ_S with

$$\Delta_R = \frac{2\sqrt{d_{avg} - 1} - \lambda_2}{\lambda_2} \tag{1}$$

$$\Delta_S = \frac{2\sqrt{\lambda_1 - 1} - \lambda_2}{\lambda_2} \tag{2}$$

We recall that these bounds were also considered in (Pal et al., 2022).

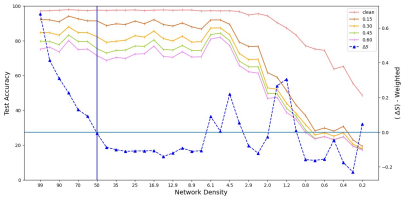
A.3 TABULAR REPRESENTATION OF THE RESULTS FROM FIGURE 3

In the table 2, k denotes the sequence length and the percentages reported are those at which the spectral bounds Δ_S and Δ_R become negative for the first time for the unweighted graph representation (the pruning mask).

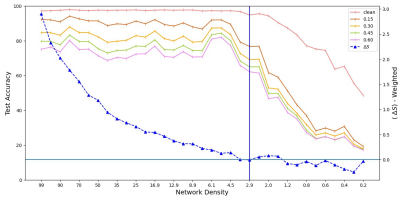
Table 2: Representation of results from Figure 3

RNN				LSTM				
Dataset	k	W_{xh} (Δ_S and Δ_R)	W_{hh} (Δ_S and Δ_R)	Dataset	k	W_{xh} (Δ_S)	W_{xh} (Δ_R)	W_{hh} (Δ_S and Δ_R)
MNIST	28	50%	12%	Speech Commands	400	50%	35%	35%

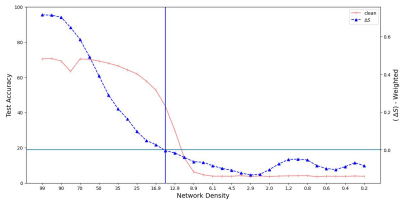
A.4 PLOTS FOR THE WEIGHTED GRAPH REPRESENTATION



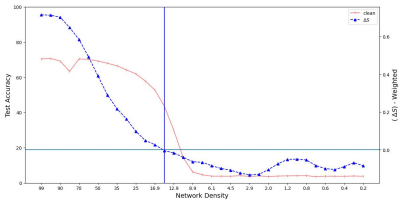
(a) W_{xh} layer of RNN (MNIST)



(b) W_{hh} layer of RNN (MNIST)



(c) W_{xh} layer of LSTM (Speech)



(d) W_{hh} layer of LSTM (Speech)

Figure 4: Variation in test set accuracy and spectral gap (Δ_S) considering **weighted** graph representation. The single vertical line shows the crossing point of both Δ_S with respect to the horizontal 0 line. In (a) and (b), plots for varying variance of zero mean gaussian noise along with clean data have been plotted.