

MinimalRNN: Toward More Interpretable and Trainable Recurrent Neural Networks

Minmin Chen

Google
Mountain view, CA 94043
minminc@google.com

Abstract

We introduce MinimalRNN, a new recurrent neural network architecture that achieves comparable performance as the popular gated RNNs with a simplified structure. It employs minimal updates within RNN, which not only leads to efficient learning and testing but more importantly better interpretability and trainability. We demonstrate that by endorsing the more restrictive update rule, MinimalRNN learns disentangled RNN states. We further examine the learning dynamics of different RNN structures using input-output Jacobians, and show that MinimalRNN is able to capture longer range dependencies than existing RNN architectures.

1 Introduction

Recurrent neural networks have been widely applied in modeling sequence data in various domains, such as language modeling [11, 9, 7], translation [1], speech recognition [4] and recommendation systems [5, 15]. Among them, Long Short-Term Memory networks (LSTM) [6] and Gated Recurrent Units (GRU) [2] are the most prominent model architectures. Despite their impressive performance, the intertwine and recurrent nature of update rules used by these networks has prevented us from gaining thorough understanding of their strengths and limitations [8].

Recent work on Chaos Free Networks (CFN) [10] inspected these popular networks from a dynamical system viewpoint and pointed out that existing RNNs, including vanilla RNNs, LSTM and GRUs, intrinsically embody irregular and unpredictable dynamics. Even without interference from input (external) data, the forward trajectories of states in these networks attract to very different points with a small perturbation of the initial state. Take GRUs as an example, it updates its states over time as follows:

$$\mathbf{h}_t = \mathbf{u}_t \odot \mathbf{h}_{t-1} + (\mathbf{1} - \mathbf{u}_t) \odot \tanh(\mathbf{W}_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{W}_x \mathbf{x}_t + \mathbf{b}_h) \quad (1)$$

where \mathbf{u}_t and \mathbf{r}_t are the update and reset gates respectively. The authors identified the multiplication $\mathbf{W}_h(\mathbf{r}_t \odot \mathbf{h}_{t-1})$ in the second part of the update, i.e., mixing of different dimensions in the hidden state, as the cause of the chaotic behavior. To address that, the authors proposed CFN, which updates its hidden states as

$$\mathbf{h}_t = \mathbf{u}_t \odot \tanh(\mathbf{h}_{t-1}) + \mathbf{i}_t \odot \tanh(\mathbf{W}_x \mathbf{x}_t + \mathbf{b}_x). \quad (2)$$

Here \mathbf{u}_t and \mathbf{i}_t are the update and input gates. By ruling out the mixing effect between the different dimensions in the hidden state, the network presents a much more predictable dynamic. The simpler network achieves comparable performance as the more dynamically complex LSTMs or GRUs for various language modeling tasks.

Inspired by the success of CFN, we propose another recurrent neural network architecture named Minimal Recurrent Neural Networks (MinimalRNN), which adopts minimum number of operations

within RNN without sacrificing performance. Simplicity not only brings efficiency, but also interpretability and trainability. There have been evidences that favorable learning dynamic in deep feed-forward networks arises from input-output Jacobians whose singular values are $O(1)$ [14, 13]. We empirically study the input-output Jacobians in the scope of recurrent neural networks and show that MinimalRNN is more trainable than existing models. It is able to propagate information back to steps further back in history, that is, capturing longer term dependency.

2 Method

Figure 1 illustrates the new model architecture named MinimalRNN. It trades the complexity of the recurrent neural network with having a small network outside to embed the inputs and take minimal operations within the recurrent part of the model.

At each step t , the model first maps its input \mathbf{x}_t to a latent space through

$$\mathbf{z}_t = \Phi(\mathbf{x}_t)$$

$\Phi(\cdot)$ here can be any highly flexible functions such as neural networks. In our experiment, we take $\Phi(\cdot)$ as a fully connected layer with tanh activation. That is, $\Phi(\mathbf{x}_t) = \tanh(\mathbf{W}_x \mathbf{x}_t + \mathbf{b}_z)$.

Given the latent representation \mathbf{z}_t of the input, MinimalRNN then updates its states simply as:

$$\mathbf{h}_t = \mathbf{u}_t \odot \mathbf{h}_{t-1} + (\mathbf{1} - \mathbf{u}_t) \odot \mathbf{z}_t \quad (3)$$

where $\mathbf{u}_t = \sigma(\mathbf{U}_h \mathbf{h}_{t-1} + \mathbf{U}_z \mathbf{z}_t + \mathbf{b}_u)$ is the update gate.

Latent representation. The dynamic of MinimalRNN prescribed by these updates is fairly straight-forward. First, the encoder $\Phi(\cdot)$ defines the latent space. The recurrent part of the model is then confined to move within this latent space. At each step t , MinimalRNN takes its previous state \mathbf{h}_{t-1} and the encoded input \mathbf{z}_t , then simply outputs a weighted average of both depending on the gate \mathbf{u}_t . That is, dimension i of the RNN state h_t^i is activated by input z_t^i and relax toward zero without any new input from that dimension. The rate of relaxation is determined by the gate u_t^i . It gets reactivated once it sees z_t^i again.

Comparing with LSTM, GRU or CFN, MinimalRNN resorts to a much simpler update inside the recurrent neural network. It retains the gating mechanism, which is known to be critical to preserve long range dependencies in RNNs. However, only one gate remains. The update rule bears some similarity to that of CFN, in that both forbid the mixing between different dimensions of the state.

Trainability. Recurrent neural networks are notoriously hard to train due to gradient explosion and vanishing [12, 3]. Several recent works [14, 16, 13] study information propagation in deep networks and suggest that well-conditioned input-output Jacobians leads to desirable learning dynamic in deep neural networks. In particular, if every singular value of the input-output Jacobians remains close to 1 during learning, then any error vector will preserve its norm back-propagating through the network. As a result, the gradient will neither explode nor vanishing. Thanks to the simple update rule employed in MinimalRNN, we can easily write out the input-output Jacobian, i.e., derivatives of RNN state \mathbf{h}_t w.r.t. input \mathbf{x}_{t-k} as follows:

$$\frac{\partial \mathbf{h}_t}{\partial \mathbf{x}_{t-k}} = \left(\prod_{t-k < i \leq t} \frac{\partial \mathbf{h}_i}{\partial \mathbf{h}_{i-1}} \right) \frac{\partial \mathbf{h}_{t-k}}{\partial \mathbf{z}_{t-k}} \frac{\partial \mathbf{z}_{t-k}}{\partial \mathbf{x}_{t-k}} \quad (4)$$

$$\text{where } \frac{\partial \mathbf{h}_i}{\partial \mathbf{h}_{i-1}} = D_{\mathbf{u}_i} + D_{(\mathbf{h}_{i-1} - \mathbf{z}_i) \odot \mathbf{u}_i \odot (\mathbf{1} - \mathbf{u}_i)} \mathbf{U}_h$$

Here $D_{\mathbf{u}}$ denotes a diagonal matrix with vector \mathbf{u} as the diagonal entries. Assuming the weight matrix \mathbf{U}_h is unitary, that is, the singular values of \mathbf{U}_h are all 1, then we can easily see that the maximum singular value of $\frac{\partial \mathbf{h}_i}{\partial \mathbf{h}_{i-1}}$ is bounded by 1. Similarly for $\frac{\partial \mathbf{h}_{t-k}}{\partial \mathbf{z}_{t-k}}$.

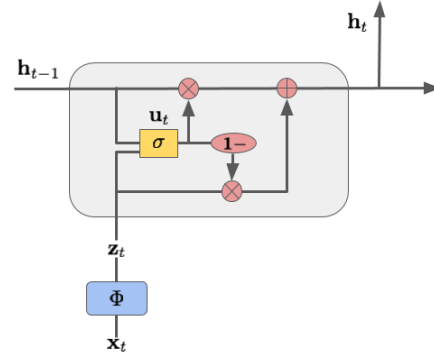


Figure 1: Model architecture of MinimalRNN.

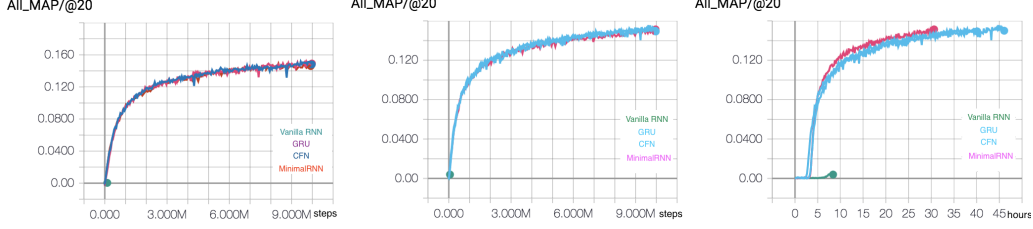


Figure 2: MAP@20 evaluated on the test sets progressed over 10M learning steps.

In comparison, the Jacobian of GRU is much more complex,

$$\frac{\partial \mathbf{h}_i}{\partial \mathbf{h}_{i-1}} = D_{\mathbf{u}_i} + D_{(\mathbf{h}_{i-1} - \mathbf{z}'_i) \odot \mathbf{u}_i \odot (1 - \mathbf{u}_i)} \mathbf{U}_h + D_{(1 - \mathbf{u}_i) \odot (1 - \mathbf{z}'_i^2) \odot \mathbf{r}_i} \mathbf{W}_h \mathbf{R}_h \quad (5)$$

here $\mathbf{z}'_i = \tanh(\mathbf{W}_h(\mathbf{r}_t \odot \mathbf{h}_{i-1}) + \mathbf{W}_x \mathbf{x}_i + \mathbf{b}_h)$ and \mathbf{R}_h is the weight matrix used in the reset gate of GRU. The Jacobian has the additional multiplication term between \mathbf{W}_h and \mathbf{R}_h in each layer, which we hypothesize will result in GRUs more prone to exploding or vanishing gradient.

3 Experiments

We demonstrate the efficacy of our method on a recommendation task of production scale. The goal is to recommend to users items of interest given user’s historical interactions with items in the system.

Dataset. The dataset contains hundreds of millions of user records. Each one is a sequence of (itemId, pageId, time) tuples, recording the context under which a recommended item consumed by an user. We consider user activities up to several months and truncate the sequence to a maximum length of 500. The item vocabulary contains 5 million most popular items of the last 48 hours.

Setup. Our production system uses a GRU as the core mechanism to capture the evolving of user interest through time. The model takes a sequence of user actions on items, and aims to predict the next item the user is going to consume. We replace the GRU component with various recurrent neural networks architectures, such as Vanilla RNN, CFN and MinimalRNN, and compare their performance to the production system. The main performance metric that we monitor offline is the Mean-Average-Precision@20.

Performance. Figure 2 plots the MAP@20 of the recommender system with different recurrent neural networks over 10M learning steps. All the weights in the recurrent neural nets are initialized to be unitary. As our data is refreshed daily, we were able to compare these methods over multiple datasets. Figure 2 left and middle show two runs. In both cases, Vanilla RNN failed during early stage of the learning due to gradient explosion. The other three models perform rather similar, reaching MAP@20 of 0.15. Since the update in MinimalRNN is much simpler, it takes less time to train. Learning finished in 30 hours comparing with CFN which takes 36 hours and 46 hours for GRU, as shown in the right column of figure 2.

Latent representation. In this experiment, we attempt to look inside the hidden states learned by MinimalRNNs. Our intuition is that the stricter updates in MinimalRNN forces its states to reside in some latent space defined by the input encoder $\Phi(\cdot)$. Each dimension of the state space focuses on some factor of the input, but not the others.

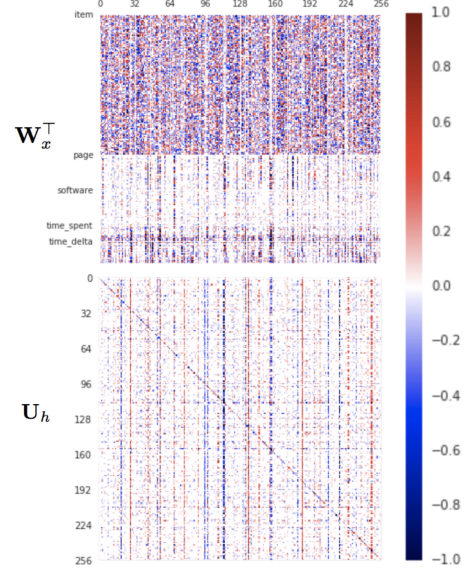


Figure 3: (Top). Weight matrix \mathbf{W}_x that transform input to latent space; (Bottom). Weight matrix \mathbf{U}_h that computes the update gate according to previous hidden state \mathbf{h}_{t-1} .

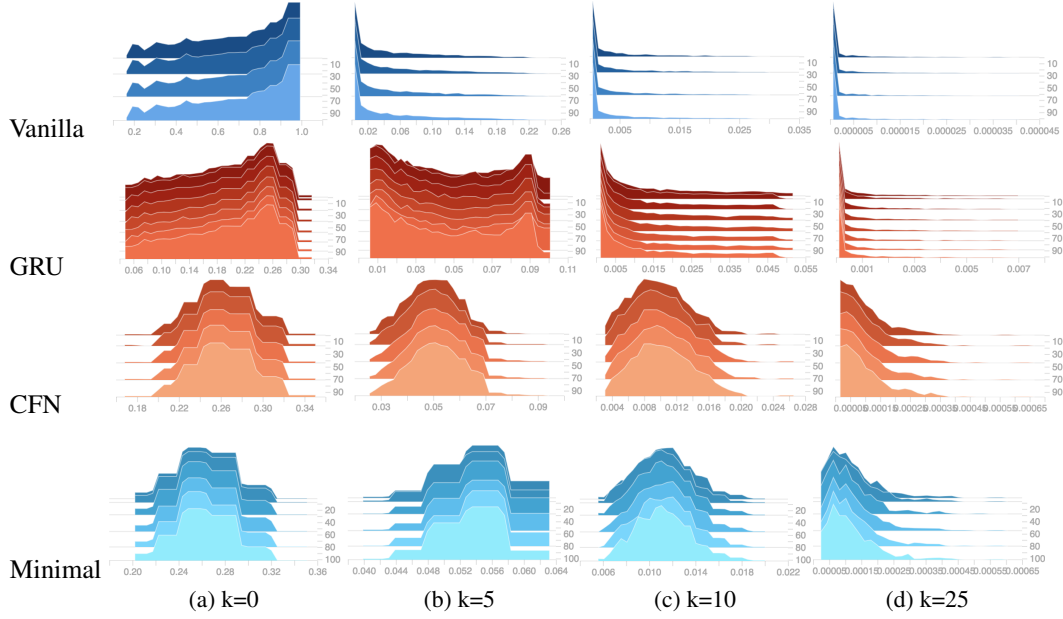


Figure 4: Histograms of the singular values of input-output Jacobians $\frac{\partial \mathbf{h}_T}{\partial \mathbf{x}_{T-k}}$ for $k = 0, 5, 10, 25$ at initial point, with weight matrices random initialized to be unitary.

The first row of figure 3 plots the weight matrix \mathbf{W}_x^\top that is used to transform the input to the latent space. Each row is one input dimension and each column is one dimension in the latent space. Entry (i, j) indicates the activation of input feature i on the latent dimension j . The input are grouped by blocks, the first block is the item embedding, and second block is the page embedding, etc.. We can see that most of the latent dimensions are used to capture the item embedding, while remaining ones capture the other contexts, such as the page on which the item is displayed, the software the user used to access the item, and time information. The bottom row of figure 3 plots the weight matrix \mathbf{U}_h that is used to compute the update gate. Each entry (i, j) indicates the activation of previous state h_{t-1}^j on the forget gate entry u_t^i . It shows several interesting properties. First, we observe strong activations on the diagonal. That is, the rate of forgetting depends mostly on the previous state from the same dimension, $h_{t-1}^i \rightarrow u_t^i$. Second, we can observe several dimensions (columns) having strong activations across all rows. In other words, these dimensions impact the rate of forgetting for almost all the dimensions in the hidden states, and these dimensions mostly corresponds to the dimensions that are capturing the context information as shown in the top of the figure.

Trainability. In these experiments, we take the input-output Jacobians computed from different recurrent neural networks at initial point and during learning to understand the trainability of these RNN structures. Figure 4 plots the histogram of the singular values of the Jacobian matrix over various k at **initial point**. All the weights in the networks are initialized to be unitary. When $k = 0$, we are looking at the derivatives of the RNN hidden state w.r.t. the current input, while $k = 25$ depicts the derivatives w.r.t. input that is 25 step back. We can see that the singular values of the input-output Jacobians in vanilla RNN quickly vanishes towards zero as k increases. The singular values of the input-output Jacobians for the GRUs starts to stretch in some directions, and shrink in others when k reaches 10, which we hypothesize is due to the additional multiplication as shown in equation (5). The input-output Jacobians of CFN and MinimalRNN on the other hand are relatively well-conditioned even for $k = 25$. Especially for MinimalRNN, the singular values stay normally distributed as k increases, and neither stretching nor shrinking in any directions.

As pointed out in [16], a good initialization does not necessarily guarantee trainability. Figure 5 plots the distribution of the singular values of the Jacobian matrices throughout the whole learning process. As learning of Vanilla RNN failed quite early, we ignore it in the comparison. We can see that the singular values of the Jacobian matrix in GRU grows rapidly in some iterations, suggesting the back-propagation error could be stretching over those directions. The singular values of the Jacobians in CFN are shrinking mostly toward 0 as learning goes on. In comparison, the Jacobians of

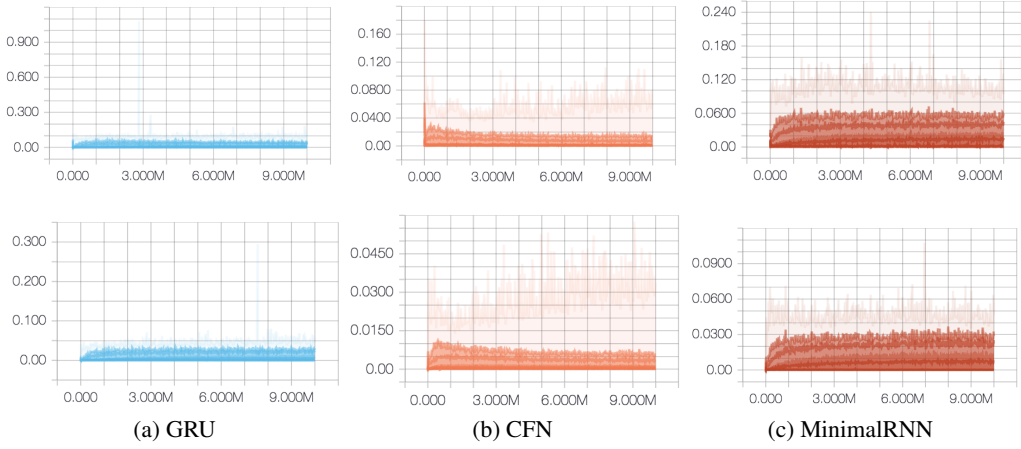


Figure 5: Distribution of singular values of input-output Jacobian $\frac{\partial \mathbf{h}_T}{\partial \mathbf{x}_{T-k}}$ at $k = 10$ (first row) and $k = 25$ (second row) during 10M learning steps. Each line on the chart represents a percentile in the distribution over the data: for example, the top line shows how the maximum value has changed over time, and the line in the middle shows how the median has changed. Reading from top to bottom, the lines have the following meaning: [maximum, 93%, 84%, 69%, 50%, 31%, 16%, 7%, minimum].

MinimalRNN are relatively well-conditioned throughout the learning. We can observe similar trends for different values of $k, k > 10$. These results suggest that MinimalRNN could be able to capture input far back in the history, i.e., longer range dependencies.

4 Future work

It remains to be seen if this extremely simple recurrent neural network architecture is able to carry over to a wide range of tasks besides the one presented here. Our most performant model for this task so far only uses one fully connected layers in $\Phi(\cdot)$. It will be interesting to find data of more complex input patterns that will require us to increase the capacity in the input encoder Φ . We would like to build upon recent success of understanding information propagation in deep networks using random matrix theory [13] to further study learning dynamics in these recurrent neural networks.

References

- [1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [2] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [3] Jasmine Collins, Jascha Sohl-Dickstein, and David Sussillo. Capacity and trainability in recurrent neural networks. *arXiv preprint arXiv:1611.09913*, 2016.
- [4] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 IEEE international conference on*, pages 6645–6649. IEEE, 2013.
- [5] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.
- [6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [7] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.
- [8] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

- [9] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.
- [10] Thomas Laurent and James H. von Brecht. A recurrent neural network without chaos. *CoRR*, abs/1612.06212, 2016.
- [11] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3, 2010.
- [12] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013.
- [13] Jeffrey Pennington, Sam Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *Advances in neural information processing systems*, 2017.
- [14] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.
- [15] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J Smola, and How Jing. Recurrent recommender networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 495–503. ACM, 2017.
- [16] Di Xie, Jiang Xiong, and Shiliang Pu. All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. *arXiv preprint arXiv:1703.01827*, 2017.