

ANALYZING INFORMATION IN RNNs

CS 294-131 Research Project

Authors:

Stefan IVANOVIC (Freshman)

email:

SID:

Benjamin KHA (Senior)

email:

SID:

Professors:

Trevor DARRELL

Dawn SONG

Vignesh MURUGANANTHAM
(Senior)

muruvig@berkeley.edu

SID: 25390657

1 Problem Definition & Motivation

In this project, we are interested in the structures that might arise in neural networks from an information theoretic perspective. More specifically, we are interested in the case of possible structures involving fully recurrent neural networks and mutual information. Our specific subdomain is within the context of using recurrent neural networks for time series prediction tasks in a continuous input and output space. There is a great deal of work done in this area showing that recurrent neural networks perform well for this task, for example in [1] and [2]. Even more specifically, here are 5 questions we will research in exploring this area:

1. Is there a general direction of information flow within the Neural Network?
2. Does the network naturally form structures similar to layers?
3. Does the concept of an Information Bottleneck apply in a useful way of fully recurrent neural networks?
4. Do features such as complexity and mutual information with Y give useful indicators of learning?
5. Do results from feed forward neural networks about the drift phase and diffusion phase also apply to fully recurrent neural networks?

2 Related Work & Comparisons

Two important pieces of related work are “Opening the black box of Deep Neural Networks via Information” by Ravid Schwartz-Ziv and Naftali Tishby [3], and “Information Theory for Analyzing Neural Networks” by Bård Sørngård [4]. These papers present some foundational work on using information theory in deep learning research.

The results in “Opening the black box of Deep Neural Networks via Information” are perfectly sufficient for analyzing feed forward neural networks, however, there is no analysis of recurrent neural networks. “Information Theory for Analyzing Neural Networks” does analyze recurrent neural networks, however, it does not consider the topics we wish to analyze. These topics include fully recurrent neural networks, analyzing the development of structure

in neural networks (this RNN has a very simple, non-flexible predetermined structure), the stages of learning in neural networks, and the concept of an information bottleneck.

3 Approach

We started by using an existing implementation of an RNN used to predict future temperatures for a certain airport given the temperature in the past. Thus we have a continuous input and output space for this problem. This was taken from [5]. This luckily came with data files containing the results of their tests given as a .pkl file. Make references to [6] and [7].

4 Current Progress

5 Timeline

1. Completing a working fully recurrent neural network for our problem. - 3/10
2. Completing a method of analyzing information transfer between neurons, and the mutual information between neurons (and groups of neurons) and X or Y. - 3/17
3. Creating a directed graph of the network, analyzing the graph to understand the structure of the network, and analyzing the direction of information transfer in the network. -3/20
4. Analyze the relationships between complexity, error, and mutual information with Y for our network. 3/28
5. Complete analyses of how the mutual information with X and Y change during training. Also analyze how the gradient mean and standard deviation behave during training. 4/4
6. Use our analyses to conclude how the concept of an information bottleneck applies to fully recurrent neural networks. Also use our analyses to

conclude if the network’s training also splits into a drift phase and diffusion phase. If not, analyze if the networks training has any apparent phases (not necessarily these two). 4/11

7. Form conclusions about how all of these analyses relate to each other, what results seem the most important, and what would be interesting for other researchers to further look into. 4/18
8. Complete our paper with information on all important results. 4/23

6 References

References

- [1] J. T. Connor, R. D. Martin, and L. E. Atlas. Recurrent neural networks and robust time series prediction. *IEEE Transactions on Neural Networks*, 5(2):240–254, Mar 1994.
- [2] E. W. Saad, D. V. Prokhorov, and D. C. Wunsch. Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks. *IEEE Transactions on Neural Networks*, 9(6):1456–1470, Nov 1998.
- [3] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017.
- [4] Bård Sørngård. Information theory for analyzing neural networks. Master’s thesis, Norwegian University of Science and Technology, 2014.
- [5] Mohammad Reza Khalghani Mohamad Aljazaery, Ahmed Cheikh Sidiya. Final report: Weather forecast. Unpublished graduate school paper, 2016.
- [6] Alessandro Achille and Stefano Soatto. On the emergence of invariance and disentangling in deep representations. *CoRR*, abs/1706.01350, 2017.
- [7] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. *CoRR*, abs/1503.02406, 2015.