Analyzing Information in RNNs

CS 294-131 Research Project

Authors:
Stefan Ivanovic (Freshman)
email:
stefanivanovic@berkeley.edu
SID:

Benjamin Kha (Senior)
email: ben.kha@berkeley.edu
SID: 25290527

Vignesh Muruganantham
(Senior)
email: muruvig@berkeley.edu
SID: 25390657

Professors:
Trevor Darrell
Dawn Song

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. Most of the previous work involving recurrent neural networks are concerned with different applications, usually natural language [**?**] or speech recognition [**?**], but there has also been some work involving time series as well [**?**, **?**] and in some other continuous input and output space settings [**?**]. Even more specifically, here are 5 questions we will research in exploring this area as our problem definitions:

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

We were inspired by the presented work in the class on information theory [**?**], making us research this topic more in depth. Two important pieces of related work are "Opening the black box of Deep Neural Networks via Information" by Ravid Schwartz-Ziv and Naftali Tishby [**?**], and "Information Theory for Analyzing Neural Networks" by Bård Sørngård [**?**]. 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, similar to that discussed in Tishby & Zaslavsky's paper [**?**].

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 [**?**]. This luckily came with data files containing the results of their test set given as a .pkl file. Using this, we measured the time delayed mutual information in the network to see if there was any clear pattern (this was done using a mutual information estimator for continuous variables [**?**, **?**, **?**]).

4 Current Progress

So far, we have conducted the analysis we wanted on the given pickled test set data. We programmed code to save all the relevant data during testing to be used in our analyses. We programmed an efficient way of processing the data and calculating the time delayed mutual information between every neuron in the network (including inputs, outputs, and hidden neurons). We calculated the entropy of each neuron. We also calculated non-time-delayed mutual information between the neurons and subtracted this from the time delayed mutual information. This had the effect of removing the effects of confounding variables and allowing us to see the information flow between neurons more effectively. We then looked to find a pattern for information flow as one would expect in an RNN. We ended up using two methods. The first method calculates the progression of a neuron along the network using a directed graph of information flow, and the second method uses the raw information flow data. We found that generally the outputs were

considered "further along" the network than the hidden neurons, which were considered "further along" the network than the inputs. We then created a histogram of the progression along the network for the inputs, outputs, and hidden neurons. Using these results, it seems there is a general direction of information flow within the network. However, due to the chaotic structure of the network, and the only slight existence of a direction of information flow, we may need to perform more analysis and are not yet sure if the RNNs form structures similar to layers. We next looked into the topic of information bottleneck and found that there was a slight positive correlation between the hidden neurons and the inputs and outputs, which would be strange if the information bottleneck effect were in play. One explanation is that some neurons are simply "more important" and take into account the most significant pattern between the inputs and outputs, but more analysis is needed here. We also ran the entire training process on our own machines and are moving towards performing similar analyses on this data set.

5 Timeline

1. 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. Also finish the analysis of the training data. 4/4

2. 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

3. 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

4. Complete our paper with information on all important results. 4/23