Computer Science Tripos: Part II Project Proposal

Measuring mutual information within Neural networks

Andrius Grabauskas, ag939
Robinson College
date

Project Originator:	Andrius Grabauskas	
Project Supervisor:	Dr. Damon Wischik	**Signature:**
Director of Studies:	Prof. Alan Mycroft	**Signature:**
Overseers:	Dr. Robert Mullins	**Signature:**
	Prof. Pietro Lio'	**Signature:**

Introduction and Description of the Work

The goal of this project is to confirm or deny the results produced by Shwartz-ziv & Tishby in their paper "Opening the black box of Deep Neural Networks via Information"[1]

The paper tackles our understating of Deep Neural Networks (DNNs). As of yet there is no comprehensive theoretical understanding of how DNNs learn from data. The authors proposed to measure how information travels within the DNNs layers.

They found that training of neural networks can be split into to two distinct phases: memorization followed by the compression phase.

- memorization - each layer increases information about the input and the label
- compression - this is the generalization stage where each layer tries to forget details about the input while still increasing mutual information with the label thus improving performance of the DNN. This phase takes the wast majority of the training time.

They found that each layer in neural network tries to throw out unnecessary data from the input while preserving information about the output/label. As the network is trained each layer preserves

[1] https://arxiv.org/abs/1703.00810

more information about the label

The results they found were interesting but also contentious as they have not yet provided a formal proof, just experimental data as a result there are many peers that are cautious and sceptical of the theory.

Starting Point

I have watched a talk that Prof. Tishby gave on this topic at Yandex, no other preparation was done.

Resources Required

I will use my own laptop to train and evaluate the neural networks as I expect them to be small given the difficulty of measuring mutual information and entropy in neural networks.

For backups I intend to store my work on GitHub and my own personal machine. In case my laptop breaks I will get another one or use the MCS machines.

Substance and Structure of the Project

The aim of this project to reproduce the results provided by Prof. Tishby and his colleagues. The intention of my work is to help settle the debate surrounding the topic either strengthening the arguments in favour of the theory in case my results are inline with the aforementioned results or encourage discussion in case my results contradict the theory.

My work will require me to learn a great deal of Information theory and DNN theory. In order for my results to be valid.

One of the more contentious parts of my project will be measuring mutual information between the input a layer in the DNN and the label. Mutual information is known to be tricky to measure when it comes to Continuous random variables.

Will need to use Python to train the neural networks and GNUplot or alternative to plot the results.

Success Criteria

Confirm or deny the results produced in“Opening the black box of Deep Neural Networks via Information“ paper on the same dataset as the paper. In order to do that I will need to:

- Train a neural network on the same dataset that is used in the paper. I will need to train the DNN multiple times in order to make sure the results hold irrespective of the neuron weights. I expect a single training round to be quite light on the computational resources required but since I will need to retrain it multiple times I might need to use compute provided by the Computer Lab.

- Measure the mutual information between input the layers and the label. I will need to measure how information travels between layers this might be tricky because as I understand measuring information is know to be difficult to get right.
- And lastly I will need to analyze the results, compare them to the papers finding and address any discrepancies found.

Extensions

Provided I achieve the success criteria I am planning to explore other datasets and try to verify that the theory holds. The dataset used in the original paper is very non-standard so it would be very interesting to see how the theory holds up when we use it with different datasets.

Different datasets will most likely need different ways to approximate mutual information between layers.

Schedule

- **15th Oct – 4th Nov**

  At the start of my dissertation I expect to spend quite a bit of time reading up on the relevant subjects

  - Information theory primarily will read Mackay's book[2]
  - Information bottleneck will use relevant papers and talks.
  - Training neural networks.
  - Measuring Entropy and Mutual information between DNN layers

- **5th Nov – 18th Nov**

  At this point I should be confident enough with the theory proposed by Tishby and will start examining the code that was provided[3]. I will need to learn the relevant technologies for training neural networks.

- **19th Nov – 2nd Dec**

  Having examined Tishby's code I will reimplement it both in order to understand it better and for the cause of independent verification.

- **3rd Dec – 6th Jan**

  Having a working system to test data sets I will try to reproduce results from the paper on the same dataset. This will achieve my success criteria.

- **7th Jan – 17th Feb**

  Assuming everything goes as planned I will start looking into other datasets to verify the theory still stands and is not just a consequence of the dataset chosen in the paper.

[2]Informaition Theory, Inference, and Learning Algorithms by David J. C. MacKay
[3]https://github.com/ravidziv/IDNNs

- **18^{th} Feb – 17^{th} Mar**

  Will use the remaining time to write up the dissertation.