Andrius Grabauskas

Measuring mutual information in Neural Networks

Computer Science Tripos – Part II

Robinson College

Monday 8th April, 2019

Proforma

Name:	**Andrius Grabauskas**
College:	**Robinson College**
Project Title:	**Measuring mutual information in Neural Networks**
Examination:	**Computer Science Tripos – Part II, July 2001**
Word Count:	**????**[1]
Project Originator:	Dr. Damon Wischik
Supervisor:	Prof. Alan Mycroft

Original Aims of the Project

Work Completed

Special Difficulties

[1] This word count was computed by `detex diss.tex | tr -cd '0-9A-Za-z \n' | wc -w`

Declaration

I, Andrius Grabauskas of Robinson College, being a candidate for Part II of the Computer Science Tripos, hereby declare that this dissertation and the work described in it are my own work, unaided except as may be specified below, and that the dissertation does not contain material that has already been used to any substantial extent for a comparable purpose.

Signed

Date

Contents

List of Figures

Acknowledgements

Chapter 1

Introduction

This is the introduction

Chapter 2

Preparation

Before developing a plan for how we are going to realize the project in code we needed to fully understand the ideas presented in the paper:

- We needed to identify the main ideas of the paper and understand why some parts of the paper are not agreed upon in the scientific community. Understand why his ideas are contentious and whether reproducing his experiments could bring more validity to his claims. This involved reading papers published by Tishby and academics who shown an opposing view to him.

- A main tool that the paper relies on is MIE (Mutual Information Estimation). Reading about MIE we quickly understood that MIE is a contentious part of the project as a result we had to do a decent amount of research regarding the subject. MIE is difficult because we are trying to estimate information between two continuous distributions using only a discrete sample set. This area has not seen much academic attention so the tools we ended up using could be greatly improved in the future.

Once we had a reasonable understanding of the ideas in the paper and which areas needed more attention we diverted our attention to figuring out the details of how the experiments were conducted figure out what hyper parameters Tishby decided are important and what assumptions he made whilst devising the experiments.

In addition we needed to find out what resources are available to us online, what programming frameworks we are going to use for the projects implementation, and to think about possible extensions to the project once the success criteria has been achieved.

- Online Resources: The two main papers by Tishby and by Saxe have made their code public online via Github, we made

  Online Resources: The two main papers we were looking at has made their code available to the public via Github, the papers are Tishby‘s paper and the main opposing paper by Saxe.

- Programming frameworks: The original experiment implementation by Tishby has used the Tensorflow framework. We have decided to use the Keras framework as it produces code that is more concise and is easier to read/maintain. Furthermore

rewriting the experiments in a different framework means that we cannot rely on the details of Tishby‘s and potentially avoid any mistakes that may exist in the original implementation.

- theoretical basis
  - MIE, what Tishby has actually shown and why is it contentious
  - Had to fully understand the main points Tishby was making and other peoples opinions in the field in order to avoid downfalls that might have already been explored.
  - Had to do research MIE‘s (Mutual Information Estimator‘s) from the start we knew that MIE will be a contentious part of the project as there has been little work done (to the best of my knowledge) on estimating mutual information of continues distributions from discrete sample sets
- I needed to come up with a plan on how am I gonna reproduce the results
  - understand the setup up of Tishby‘s experiments exactly how he produced the results, thankfully the code was publicly available Tishby recently released his results on Github which I used as a point of reference.
  - figure out what technologies I am going to use. For his implementation Tishby used the Tensorflow framework. I’ve decided to use the Keras framework as I’ve found out Keras produced code that is more concise and easier to read/maintain and easy to read. Further more rewriting the code in a different framework made so that I can’t rely on the detail of Tishby‘s code and avoid mistakes that might be in the original code.
- starting point
  - Had Tishby‘s paper and code as a reference point of what has to be achieved
  - ended up following other’s peoples implementations of MIE‘s closely some of which didn’t work and had to be scrapped.

Preparation for the project required me to understand how Tishby arrived to the conclusion that he has presented us in his paper, I needed to understand how the experiments in the paper were constructed in order to be able to reproduce them reliably.

I needed to understand

Finally In order to be able to reproduce the results and extend on the ideas I needed to learn python and frameworks such as Keras and Tensorflow.

- Tensorflow - this is the framework Tishby decided to use for his implementation.
- Keras - this is the framework I have decided to use for my implementation as the produced code is much more concise.

Chapter 3

Implementation

Chapter 4

Evaluation

Chapter 5

Conclusion

Appendix A

Project Proposal

Computer Science Tripos: Part II Project Proposal

Measuring mutual information within Neural networks

Andrius Grabauskas, ag939
Robinson College
Saturday 20th October, 2018

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

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 (DNN's). As of yet there is no comprehensive theoretical understanding of how DNN's learn from data. The authors proposed to measure how information travels within the DNN's 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 more information about the label

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

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 even a paper[2] was produced that tries to suggest that the theory is wrong, however this was dismissed by Tishby & Shwartz-Ziv[3]

Starting Point

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

Resources Required

The training DNN's and measuring mutual information will be computationally expensive so I will be using Azure cloud GPU service to acquire the required compute for this project. The GPU credits will be provided by Damon Wischik

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 have a comprehensive understanding of Information theory, Information bottleneck and neural networks.

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. It will be computationally expensive to measure it in DNN since we will need to retrain the network in order to get a distribution rather than a single value. I will use Gaussian approximation to measure it (relevant paper[4])

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

Success Criteria

Reimplement the code that was used to generate the papers results. 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

[2]https://openreview.net/pdf?id=ry_WPG-A-
[3]https://openreview.net/forum?id=ry_WPG-A-&noteId=S1lBxcE1z
[4]https://arxiv.org/abs/1508.00536

that was used in the paper and measure mutual information between the layers. Analyse the results produced and address any discrepancies that may have occurred.

Extensions

Provided I achieve the success criteria there are two main ways to extend it.

- Use different datasets to test the theory. Using different datasets would confirm that the results are not data specific. Current datasets we are considering: MNIST[5] and NOT-MNIST[6].
- Explore different ways of measuring mutual information. One interesting way would be to explore a discrete neural network where every node would only be able assigned discrete values say 1...256. This would make the distribution within a DNN layer discrete and hence it would make calculating mutual information straightforward. However quantizing the neural network could possibly hurt the performance of the network.

[5]http://yann.lecun.com/exdb/mnist/
[6]https://www.kaggle.com/quanbk/notmnist

Schedule

- **20th Oct – 2nd Nov**

  I expect to spend the first two weeks reading up on Information theory (primarily from Mackay's book[7]) and the information bottleneck method in order to understand the nuances of the paper.

- **3rd Nov – 30th Nov**

  The following weeks I intend to spend reading up on DNN's doing some introductory courses, I will train the neural network on the same data as the paper but at this point will not yet try to measure the mutual information between the layers.

  At this point I will also start examining the code[8] provided and start to implement parts of it which don't deal with information measurement.

- **1st Dec – 28th Dec**

  Will start reading up on mutual Information measurement with local Gaussian approximation.

  Implementing mutual information measurement in code.

  At this point I expect the computation to be too demanding for my machine and will need to use provided compute.

- **29th Dec – 1st Feb**

  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.

  At this point my success criteria should be completed I will spend some time writing the skeleton of the thesis. Look for any discrepancies between my results and the ones provided in the paper.

- **2nd Feb – 2nd Feb**

  Assuming everything goes as planned I will start looking into implementing one of the extensions. Which are :

  - Testing the theory on different datasets.
  - Implementing a quantized neural network implementation.

  or both, if time is in my favour.

- **3rd Feb – 2nd Mar**

  Will use the remaining time to write up the dissertation.

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