Computer Science Tripos

Part II Project Proposal Coversheet

Please fill in Part 1 of this form and attach it to the front of your Project Proposal.

Part 1

Name: CRSID:

College: Overseers: (Initials)

Title of Project:

Date of submission: Will Human Participants be used?

Project Originator: Signature: ----------

Project Supervisor: Signature: ----------

Directors of Studies: Signature: ----------

Special Resource Sponsor: Signature: ----------

Special Resource Sponsor: Signature: ----------

Above signatures to be obtained by the Student

Part 2

Overseer Signature 1: ----------

Overseer Signature 2: ----------

Overseers signatures to be obtained by Student Administration.

Overseers Notes:

Part 3

SA Date Received: SA Signature Approved:

Computer Science Tripos: Part II Project Proposal

Measuring mutual information within Neural networks

Andrius Grabauskas, ag939
Robinson College

Thursday 18th 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 (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 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 of DNNs 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 dataset specific. Current datasets we considered are 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 measuring mutual information more defined, however this would possibly hurt the performance of the whole network.

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[7]
  - 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[8]. 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.

[5] http://yann.lecun.com/exdb/mnist/
[6] https://www.kaggle.com/quanbk/notmnist
[7] Informaition Theory, Inference, and Learning Algorithms by David J. C. MacKay
[8] https://github.com/ravidziv/IDNNs

- **7^{th} Jan – 17^{th} 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.

- **18^{th} Feb – 17^{th} Mar**

  Will use the remaining time to write up the dissertation.