Measuring Mutual Information inside Deep Neural Networks

Part II progress report

Andrius Grabauskas - ag939

Saturday 9th February, 2019

On Project Background and Aims

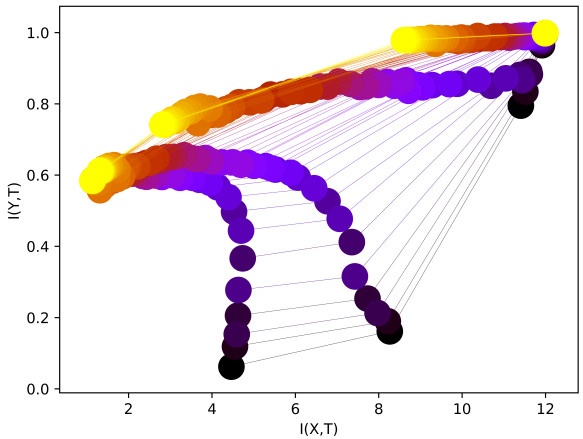
- We do not actually understand or know why DNN's learn.
- N. Tishby suggested that DNN's learn as a result of compression between the layers and the information bottleneck principle. [1]
- N. Tishby's claims have been disputed by A. M. Saxe who claimed that his results are due to the parameters he chose. [2]
- I am here to:
  - Reproduce Tishby's results provided in his paper.
  - Explore new ways to verify Tishby's core thesis.
  - Provide stable code that could be extended and used by other in the future.

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

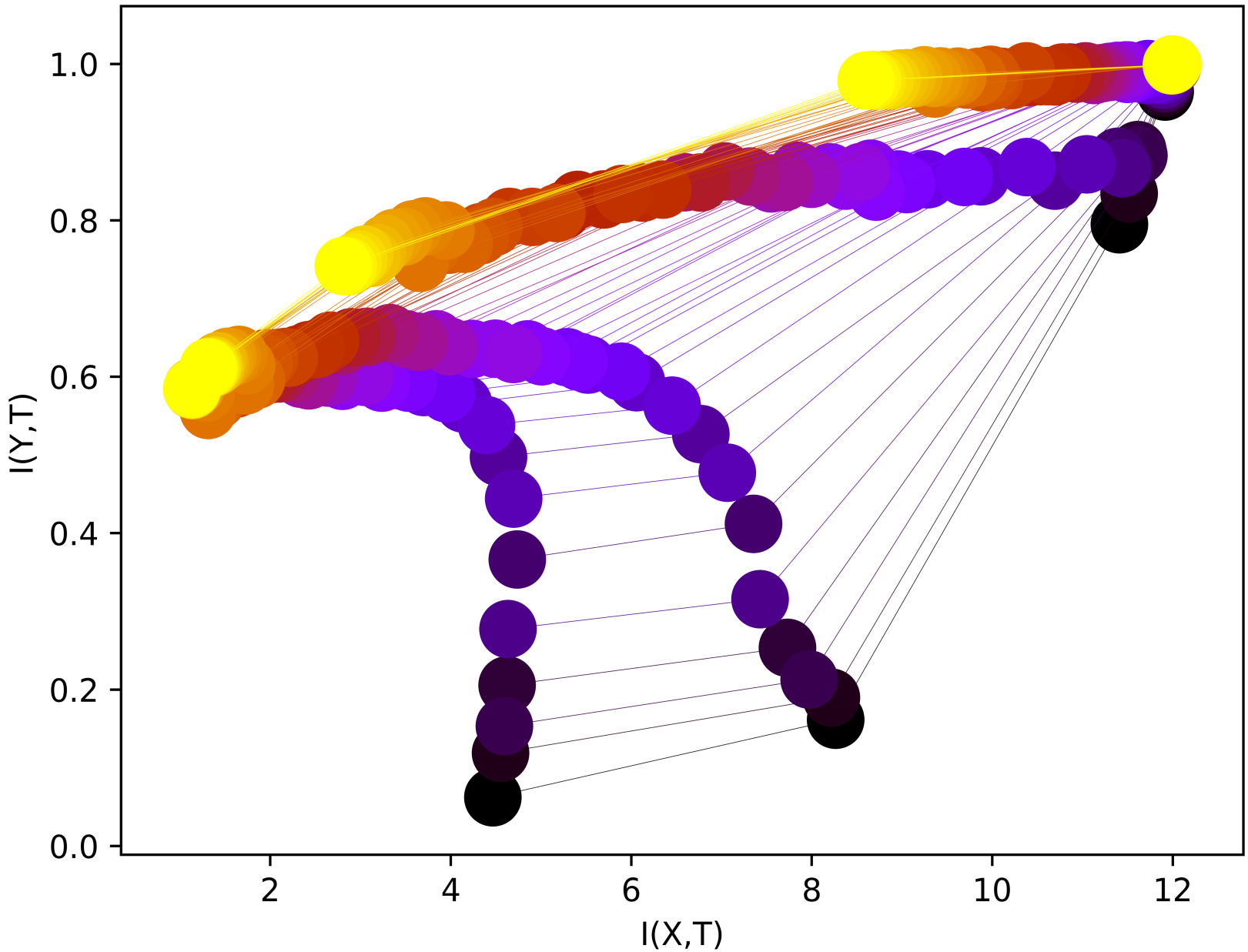
[2] https://openreview.net/forum?id=ry_WPG-A-

Success criteria

- Reproduce results in Tishby's paper
  - Essentially reproduce this image and show that the results are stable with varying parameters.



  - Each point represents a specific layer T
  - X axis represents mutual information (MI) between input data and a specific layer
  - Y axis represents mutual information (MI) between label and a specific layer
  - Each set of points connected by lines corresponds a single epoch and every layer in that epoch


1.0
0.8
0.6
0.4
0.2
0.0
I(Y,T)
2
4
6
8
10
12
I(X,T)

Current progress

- Reimplemented Tishby's code and reproduced his results.
- Found evidence that Tishby's Mutual Information Estimator (MIE) might not be completely sound.
- Had lots of trouble with finding an alternative MIE as Neural Networks are have high dimensionality which doesn't play well with some MIE.
- Settled with Kernel Density Mutual (KDE) MIE, which agreed with Tishby's results. (Although it is by far not a perfect measure of mutual information)

Remaining work, Issues

- Show that compression happens in a bigger dataset such as MNIST and hence Tishby's results holds for other datasets.
- Clean up and comment the code such that it is easy to understand and extensible by others.
- At the core of Tishby's thesis is the suggestion that weights are viewed **as if** they are random by the NN
  - However every weight is fixed at any given point in time and every weight matrix is invertible which would suggest that the neural network cannot lose information.
  - If the weights are random this implies that Neural networks truly can lose information from layer to layer.
  - Need to find ways of working around the "issue" of fixed weights.