Why Neural Networks work?

Bhavya Bhatt

Indian Institute of Technology Mandi

21 September 2019

Overview

Theoretical Deep Learning

Many approaches exist for theoretical framework of DL.
They are based on

- Information Theory
- Complexity Theory
- Statistical Computational Learning
- Group Theory

Information Bottleneck Principle

The central idea of IB-Theory is

"Regularization by optimal intermediate representations"

Mathematical Preliminary

Given $p_{XY}(x, y)$ for the dataset, IB objective is as follows

$$L(p(\hat{X}|X)) = I(\hat{X}, X) - \beta I(\hat{X}, Y)$$

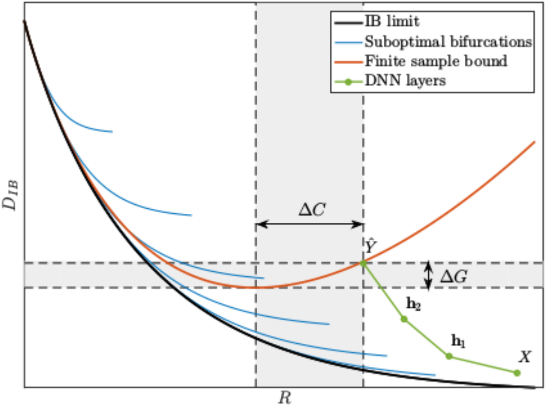
and probability distribution of minimum sufficient statistics (optimal) is

$$p^*(\hat{X}|X) = \arg\min_{p(\hat{X}|X)} L(p(\hat{X}|X))$$

So intermediate representation $\hat{X}$ is a stochastic compressed representation of X

True optimal and Empirical optimal

The optimal curve for different β with true distribution $p_{XY}(x, y)$ (in black) and with empirical distribution $\hat{p}_{XY}(x, y)$ estimated with finite samples in the dataset is as follows [1]



[1] The slope of the curve is β^{-1}

Two phase training dynamics - Generalization and Compression phase

First observed in [2]

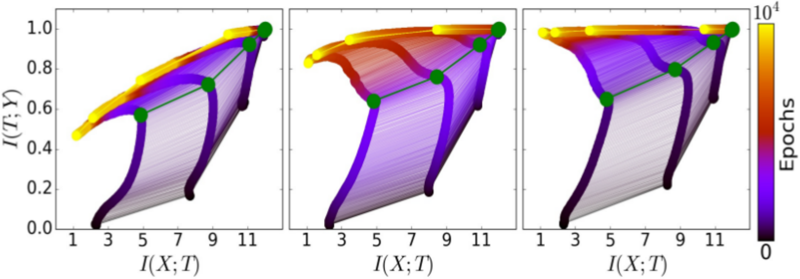


Figure: The evolution of the layers with the training epochs in the information plane, for different training samples. On the left - 5% of the data, middle - 45% of the data, and right - 85% of the data

2

[2] figure taken from [2]

The Controversy

Paper [3] attacked the original paper claiming

- IB-Theory is not fundamental theory
- Depends on specific activation used
- Showed two phase dynamics do not hold for RELU activation

But later more accurate MI estimators published and observed the two phases !

Even then , IB-Theory have issues in case of deterministic networks

Mutual Information estimation methods

Improvement in estimation methods

- Use more accurate MI estimation methods like EDGE, MINE etc. [3]
- Use tight bounded MI representations
- Research for which estimation method to use when

[3] EDGE - Ensemble Dependency Graph Estimator
MINE - Mutual Information Neural Estimator

Dependence Criterion

Instead of MI as dependence criterion use more robust criterion which captures [4]

- **inform about** Y - $\hat{X}$ should be sufficient statistics
- **be maximally compressed** - representation $\hat{X}$ should not tell about X
- **admit a simple decision function** - Y can be estimated from $\hat{X}$ using simple functions
- **be robust** - small noise should not change $\hat{X}$ with big differences

First two are captured by mutual information criterion but there is a need for criterion which captures last two too !

[4] according to the paper [5]

Use of IB-Theory

All the published work on IB-Theory can be loosely grouped into three categories:

- experimental IB-based DNN analysis (study training dynamics)
- IB-based DNN performance bounds and IB theory of DNNs
- IB-based DNN training (deep variational IB, HSIC sigma networks)

Experimental Setup

The experimental setup is as follows [5]

- dataset: MNIST item epochs: 7
- architecture: Conv2d(16, 3, 1, 1) 2 - (Flatten) - 500 - 200 - 50 - 10
- activation: RELU (for hidden layers), softmax (output layer)
- criterion: HSIC (Hilbert-Schmidt Independence Criterion)

[5] Semantics for Conv layer is (filter, kernel, stride, padding)

Results

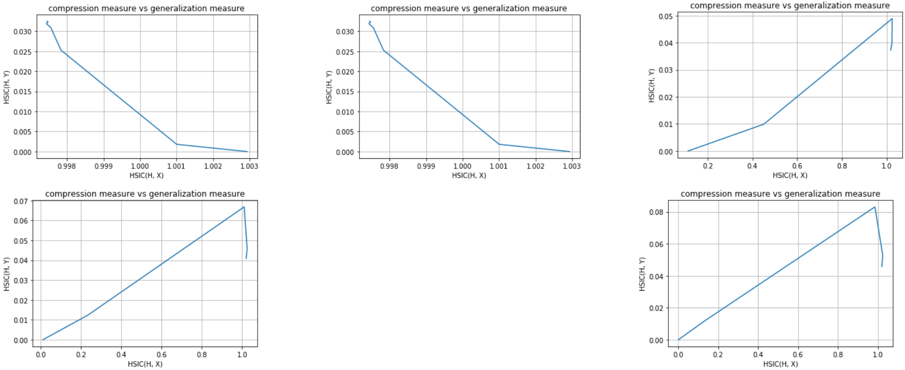


Figure: Plots between HSIC(X, H) and HSIC(H, Y) for all parametric layers - starting with first layer from top left

Results

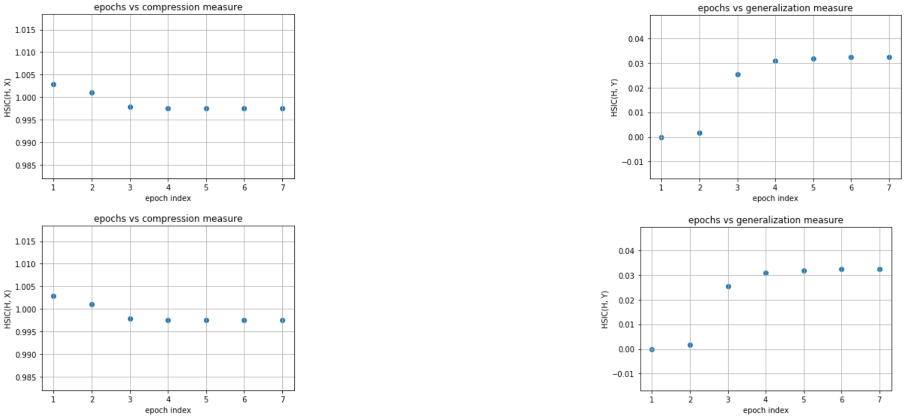


Figure: Plots between epochs vs HSIC(X, H) (left) and epochs vs HSIC(H, Y) (right) for 1st and 2nd layer

Results

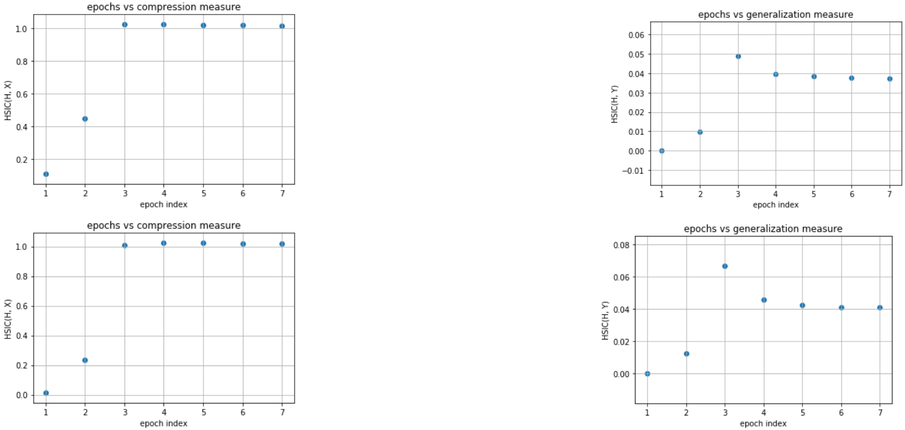


Figure: Plots between epochs vs HSIC(X, H) (left) and epochs vs HSIC(H, Y) (right) for 3rd and 4th layer

Results

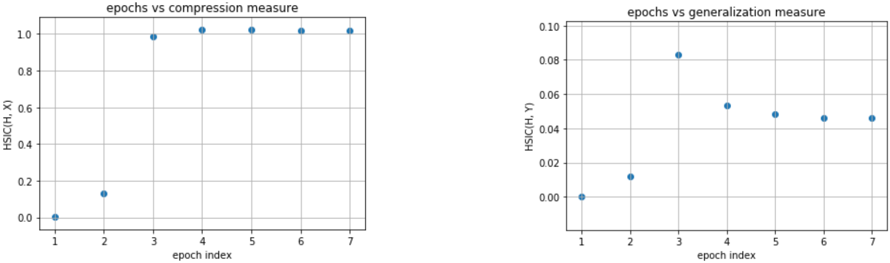


Figure: Plots between epochs vs HSIC(X, H) (left) and epochs vs HSIC(H, Y) (right) for 5th layer

Inference and Observation

As we can infer from above results

- current loss functions are insufficient in capturing generalization abilities
- network drastically over fits on the data with increasing complexity of the representation H
- decreases generalization abilities of H

The above dynamics is just opposite of what we want to achieve i.e. more generalization and less complexity for representation H

Conclusion

A theoretical framework is required which can also take into account optimal intermediate representation concept in objective for training

Scope of future work

According to the current plan the project will mainly focus on the following

- verification of IB-Theory observations
- use different MI estimation methods and different kinds of criterion
- deep variational IB models and benchmark their performance with other state-of-the-art models
- theoretically study the properties of 'optimal' intermediate representations
- explore other approaches

PyGlow: Python Package for Information Theory of Deep Learning

PyGlow

Information Theory of Deep Learning

[6]

[6] GitHub repo on: `https://github.com/spino17/PyGlow`
PyGlow Docs: `https://pyglow.github.io/`

References

- N. Tishby and N. Zaslavsky, "Deep Learning and the Information Bottleneck Principle," abs/1503.02406, 2015.
- R. Shwartz, N. Tishby, "Opening the Black Box of Deep Neural Networks via Information," arXiv/1703.00810, Aug 2018.
- A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey and D. D. Cox, "On the Information Bottleneck Theory of Deep Learning," *International Conference on Learning Representations,* 2018.
- M. Noshad and A. O. Hero III, "Scalable Mutual Information Estimation using Dependence Graphs," abs/1801.09125, Aug 2018.
- R. A. Amjad and B. C. Geiger, "How (Not) To Train Your Neural Network Using the Information Bottleneck Principle," abs/1802.09766, 2018.
- A. A. Alemi, I. Fischer, J. V. Dillon and K. Murphy, "Deep Variational Information Bottleneck," abs/1612.00410, 2016

The End