Opening the Black Box of Deep Neural Networks via Information

A paper by Ravid Schwarz-Ziv and Naftali Tishby

Jani Anttonen

17.4.2018

Department of Future Technologies
University of Turku

Outline

Background

Motivation for the Paper

- There hasn't been a clear reason why deep neural networks **are generalizing** as well as they do.
- *DNNs* don't seem to have an overfitting problem.

Motivation for the Paper

- There hasn't been a clear reason why deep neural networks **are generalizing** as well as they do.
- *DNNs* don't seem to have an overfitting problem.
- ...and there's a paper specifying a **maximal learning bound** named the **information bottleneck** for neural networks, written by the same *Tishby* as the paper presented here.

Theory – Machine Learning

- The act of distilling the **essence of information** from data (semi)automatically.

Theory – Machine Learning

- The act of distilling the **essence of information** from data (semi)automatically.
- In other words, trying to reproduce or *reverse-engineer the function* that would output the same results that exist in the data the algorithm is given.

Theory – Information Theory

- Aims to measure how much information can a **thing** (yes, everything) contain.
- Specific implementations include error correction, compression, RNGs and cryptoanalysis.

Theory – Information Theory – Mutual Information

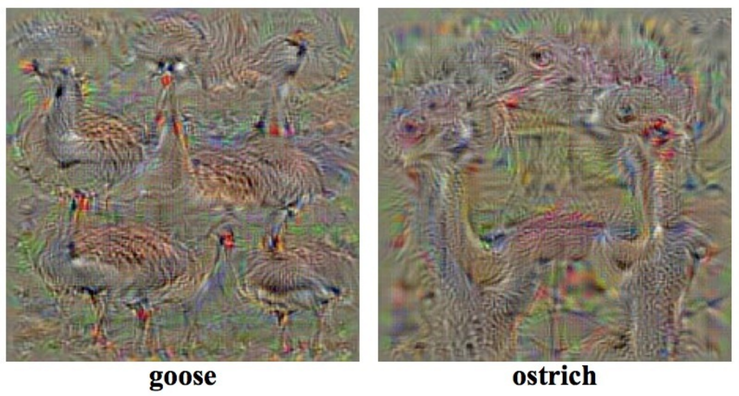
$$\mathrm{I(x,y)} = \sum_{x,y} P(x,y) \ln \frac{P(x,y)}{P(x)P(y)}$$

- Mutual information is defined with two entropies, shannon entropy and conditional entropy.
- It is a very robust *measurement of similarity*.

Looking at Neurons

- Last layer weights of a network – Closest representations of the classes
- In bigger networks, weights are somewhat recognizable. For example, in the case of Google's deep dream, they feed the network back the weights of the last neuron representing a dog, *and get snouts and drooping tongues everywhere*.

Looking at Neurons



[4]

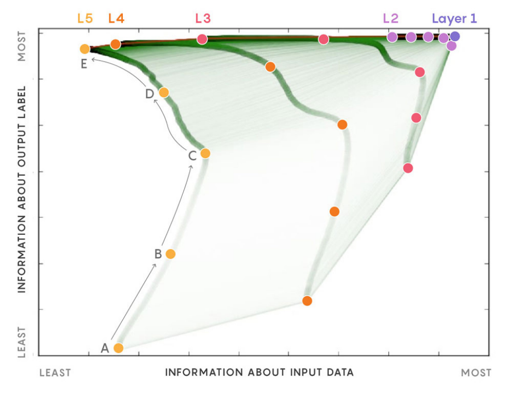
Looking at Neurons

- By iterating through a number of layers with filters, you **lose data on every one**.
- So what's basically happening is lossful compression of the whole input data!

Findings

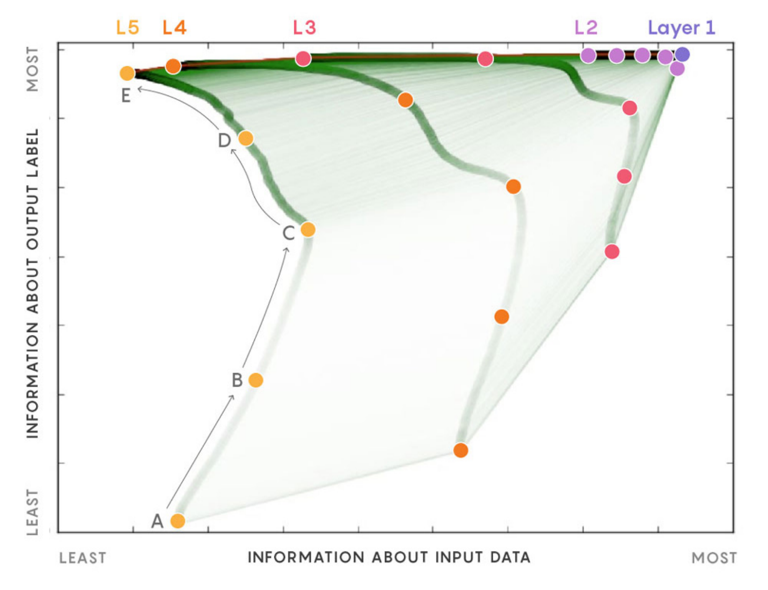
Best-case Scenario

1. Network copies the training data, gaining information of the data and the labels
2. When the gradient gets smaller and smaller it starts to slowly lose information of the original training data, compressing the representation!



[3]

Best-case Scenario



[3]

Video Time

Animation of the training progress in the information plane [2]

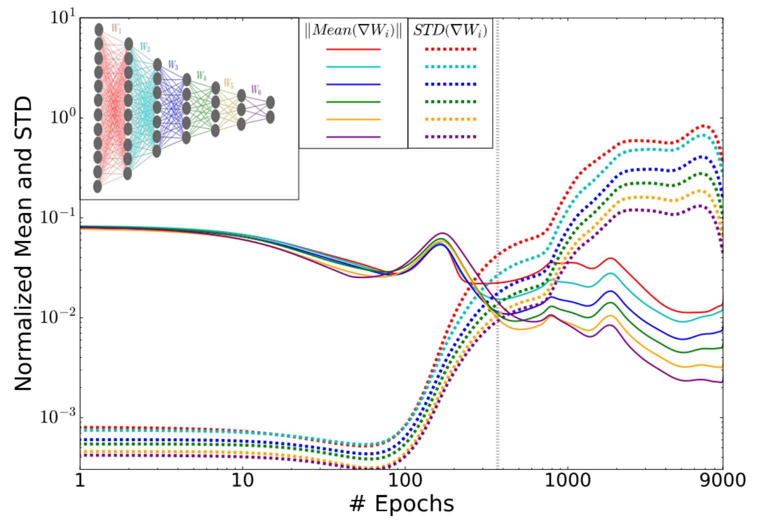
Why Does the Phase Change Happen?

- Stochastic gradient descent adds noise to the signal after a shallow gradient is reached.

Why Does the Phase Change Happen?

- Stochastic gradient descent adds noise to the signal after a shallow gradient is reached.
- This noise effectively *wipes out the irrelevant information* of the class, **compressing** the representation by relaxing the weights.

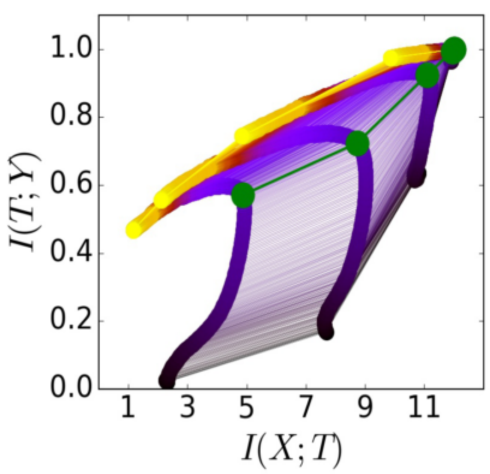
Why Does the Phase Change Happen?



[1]

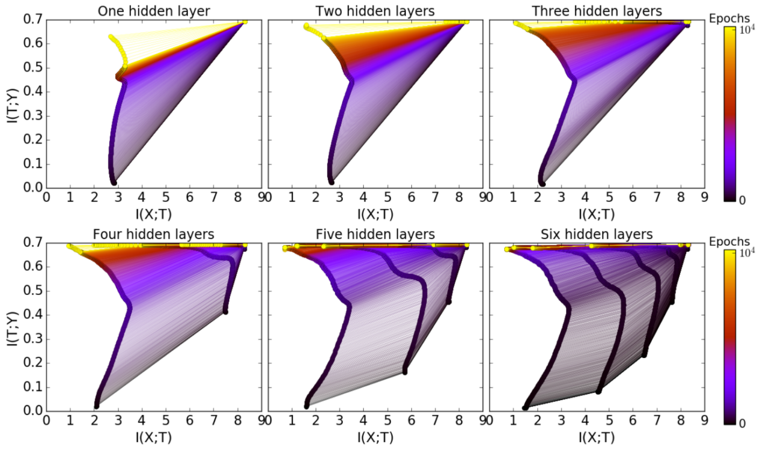
Training Gone Wrong

- The network copies the data well, but fails to generalize because it starts to lose information on the labels when compressing the representation on **shallow gradients**.
- Case: Too little data.



[1]

More Layers = Better Performance AND Speed



[1]

Summary

Summary

- The paper helps us better understand with intuition and theory, what happens during training in modern neural networks.

Summary

- The paper helps us better understand with intuition and theory, what happens during training in modern neural networks.
- This might help us to define better hyperparameters (learning rate and such) for the network beforehand.

Summary

- The paper helps us better understand with intuition and theory, what happens during training in modern neural networks.
- This might help us to define better hyperparameters (learning rate and such) for the network beforehand.
- In addition, the information bottleneck bound could be helpful in deciding if you need more data or a better network for a task.

Appendix

Sources i

Schwartz-Ziv, Ravid and Tishby, Naftali
Opening the Black Box of Deep Neural Networks via Information
arXiv:1703.00810v3, 2017.

Tishby, Naftali
Presentation: Information Theory of Deep Learning. Naftali Tishby
https://www.youtube.com/watch?v=bLqJHjXihK8, 2017.

Sources ii

Wolchover, Natalie
New Theory Cracks Open the Black Box of Deep Learning
Quanta Magazine, Wired to Learn: The Next AI 2017.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, Hod Lipson
Understanding Neural Networks Through Deep Visualization
http://yosinski.com/deepvis, 2015.