On the learning dynamics of neural nets

Sayak Paul | Deep Learning Associate at [PyImageSearch]

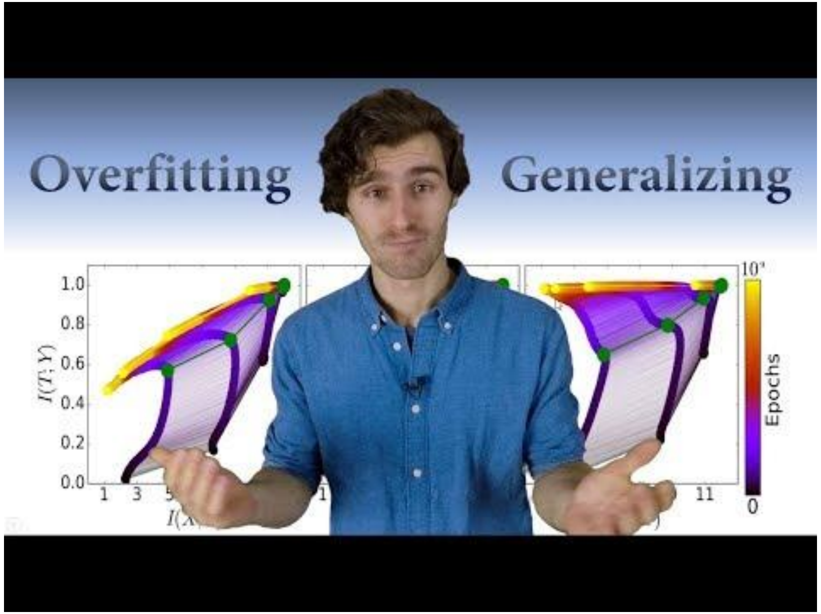
Kaggle Days Mumbai, November 30, 2019

India

Experts

Acknowledgements

- The entire team at PyImageSearch
- Xander Steenbrugge (Arxiv Insights)



sayak.dev

Agenda

- Generalization in machine learning
  - What is it?
  - Why is it important?
  - Generalization vs. Memorization: Some directions
- Deep Learning and Information Theory
- Further directions

Generalization in machine learning

- What is generalization?

Generalization in machine learning

- What is generalization?

Generalization in machine learning

- What is generalization?
  - Underfitting
  - Overfitting

Generalization in machine learning

- What is generalization?
  - Underfitting
  - Overfitting
    - Training loss is lower than validation loss ☠️

Generalization in machine learning

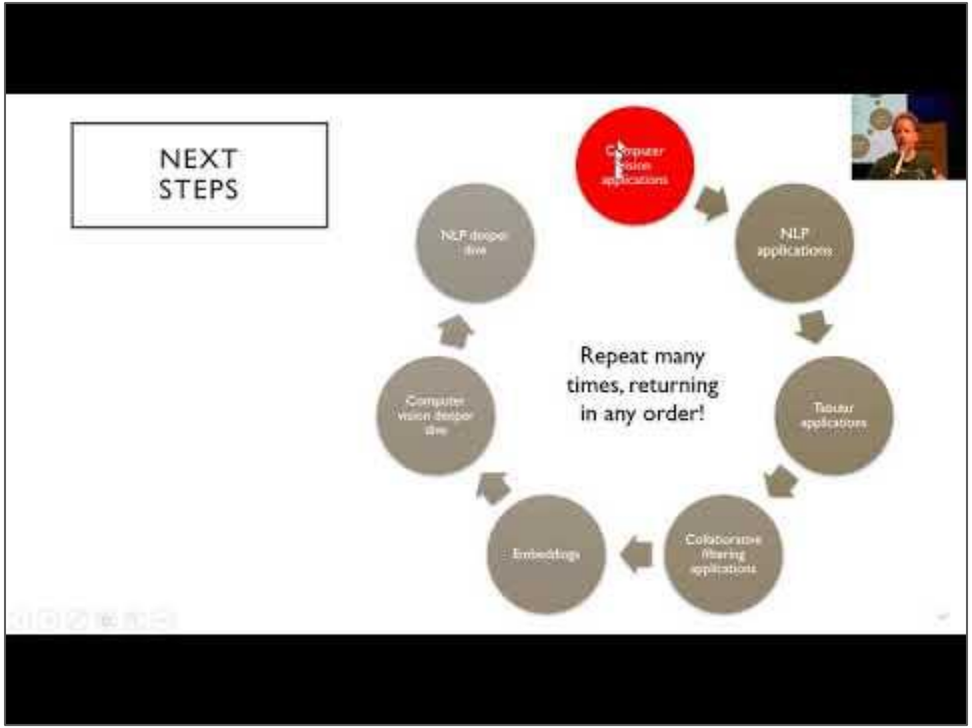
- What is generalization?
  - Underfitting
  - Overfitting
    - Training loss is lower than validation loss ☠️
    - Training accuracy is higher (much) than validation accuracy ☠️


AWARD
Global

Generalization in machine learning

- What is generalization?
  - Underfitting
  - Overfitting
    - *Overfitting is when your validation loss decrease was not steady across the epochs.*

Generalization in machine learning



[Lesson 2: Data cleaning and production; SGD from scratch]()

Generalization in machine learning

- Why is generalization important?

Generalization in machine learning

- Why is generalization important?
  - We want our models to perform equally good on the test data points.

Generalization in machine learning

- Why is generalization important?
  - We want our models to perform equally good on the test data points.
  - If a model is 50% accurate on training data points, it's generalizing if it shows similar accuracy on the test data points.

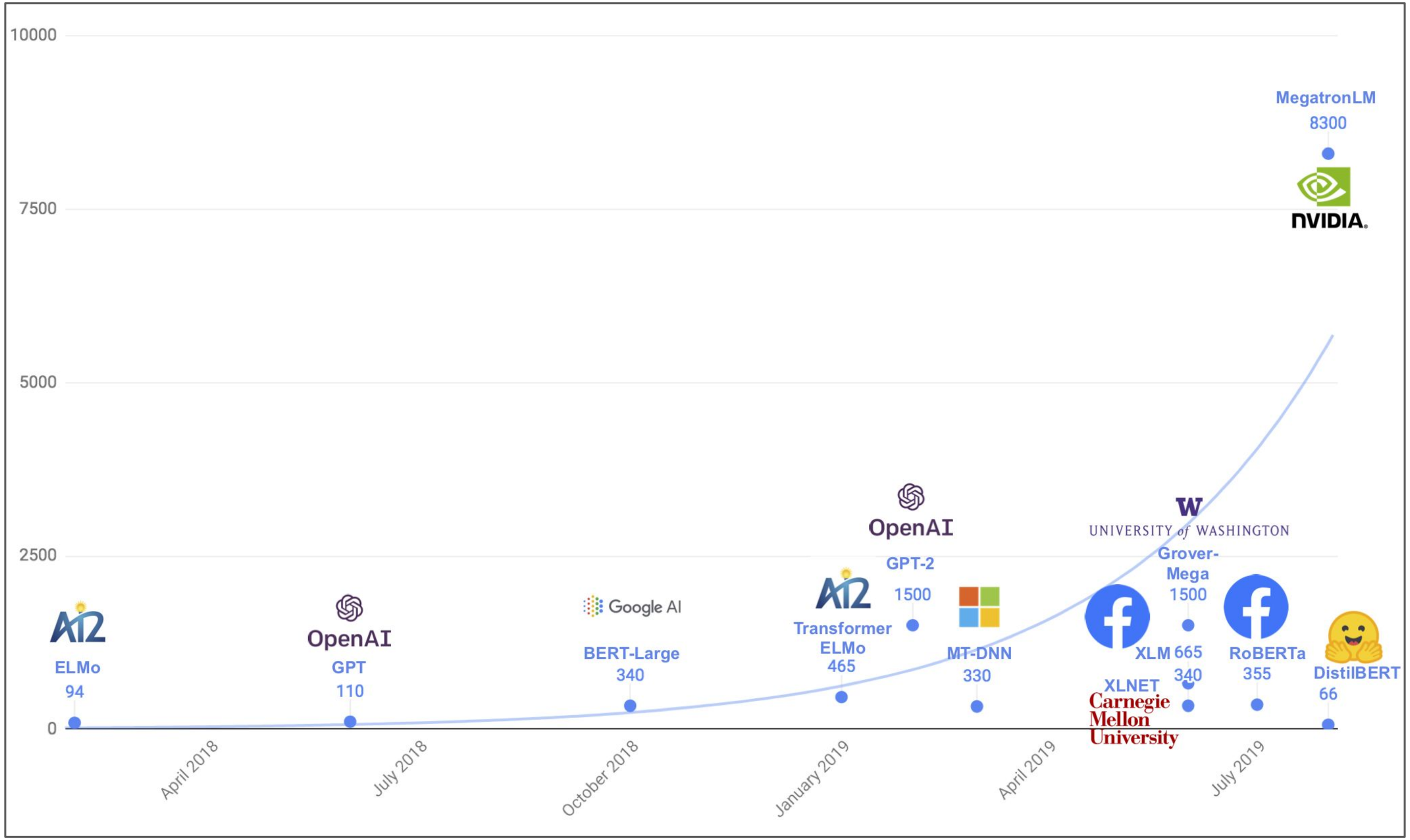
Generalization in machine learning

- Why is generalization important?
  - We want our models to perform equally good on the test data points.
  - If a model is 50% accurate on training data points, it's generalizing if it shows similar accuracy on the test data points.
  - Simple models often > very complex models

Generalization in machine learning

- Why is generalization important?
    - We want our models to perform equally good on the test data points.
    - If a model is 50% accurate on training data points, it’s generalizing if it shows similar accuracy on the test data points.
    - Simple models often > very complex models (**oh**, **really?**)

Increase in the # of parameters of DL models



sayak.dev

Generalization in machine learning

“ … the convergence of ERM is guaranteed as long as the size of the learning machine (e.g., the neural network) **does not** increase with the number of training data. “

Generalization in machine learning

“ … the convergence of ERM is guaranteed as long as the size of the learning machine (e.g., the neural network) **does not** increase with the number of training data. “

mixup: BEYOND EMPIRICAL RISK MINIMIZATION

Hongyi Zhang
MIT

Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz*
FAIR

ABSTRACT

Large deep neural networks are powerful, but exhibit undesirable behaviors such as memorization and sensitivity to adversarial examples. In this work, we propose *mixup*, a simple learning principle to alleviate these issues. In essence, *mixup* trains a neural network on convex combinations of pairs of examples and their labels. By doing so, *mixup* regularizes the neural network to favor simple linear behavior in-between training examples. Our experiments on the ImageNet-2012, CIFAR-10, CIFAR-100, Google commands and UCI datasets show that *mixup* improves the generalization of state-of-the-art neural network architectures. We also find that *mixup* reduces the memorization of corrupt labels, increases the robustness to adversarial examples, and stabilizes the training of generative adversarial networks.


THE ROOK | STARZ
WHAT IS HAPPENING HERE?

Generalization in machine learning

- What is generalization?
- Why is generalization important?
- Generalization vs. Memorization: Some directions

Generalization in machine learning

- Generalization vs. Memorization: Some directions

Generalization in machine learning

- Generalization vs. Memorization: Some directions

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

Generalization in machine learning

- Generalization vs. Memorization: Some directions
    - Crazy experiments: **Fitting random labels and pixels** 😨

Generalization in machine learning

model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75
(fitting random labels)		no	no	100.0	9.78
Inception w/o BatchNorm	1,649,402	no	yes	100.0	83.00
		no	no	100.0	82.00
(fitting random labels)		no	no	100.0	10.12
Alexnet	1,387,786	yes	yes	99.90	81.22
		yes	no	99.82	79.66
		no	yes	100.0	77.36
		no	no	100.0	76.07
(fitting random labels)		no	no	99.82	9.86
MLP 3x512	1,735,178	no	yes	100.0	53.35
		no	no	100.0	52.39
(fitting random labels)		no	no	100.0	10.48
MLP 1x512	1,209,866	no	yes	99.80	50.39
		no	no	100.0	50.51
(fitting random labels)		no	no	99.34	10.61

“Deep neural networks easily fit random labels.”

Why do the networks generalize as well?

Generalization in machine learning

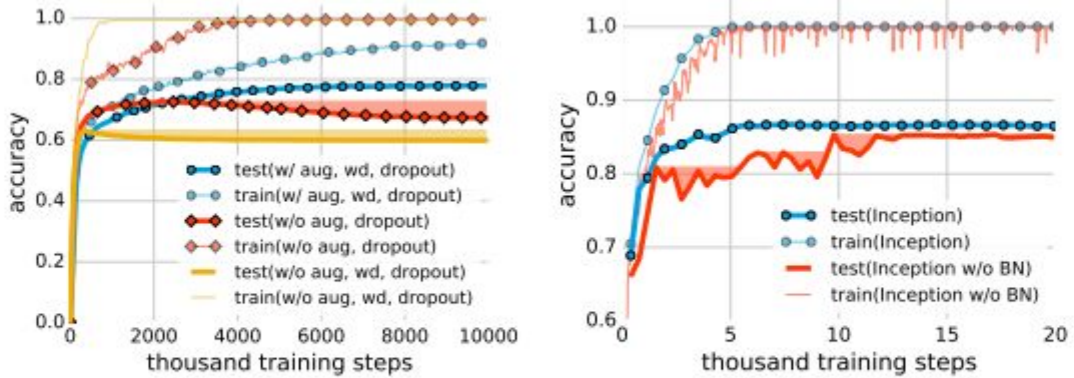
- Generalization vs. Memorization: Some directions
    - Crazy experiments: **Fitting random labels and pixels** 😨
    - **Regularization** to penalize memorization. But ...

Generalization in machine learning

- Generalization vs. Memorization: Some directions
  - Crazy experiments: **Fitting random labels and pixels** 😨
  - **Regularization** to penalize memorization. But ...
  - Studies show that even without explicit regularization networks achieve commendable performance on test data.

Generalization in machine learning

- Generalization vs. Memorization: Some directions



(a) Inception on ImageNet

(b) Inception on CIFAR10

Figure 2: Effects of implicit regularizers on generalization performance. aug is data augmentation, wd is weight decay, BN is batch normalization. The shaded areas are the cumulative best test accuracy, as an indicator of potential performance gain of early stopping. (a) early stopping could potentially improve generalization when other regularizers are absent. (b) early stopping is not necessarily helpful on CIFAR10, but batch normalization stablize the training process and improves generalization.

Generalization in machine learning

- Generalization vs. Memorization: Some directions

A Closer Look at Memorization in Deep Networks

Devansh Arpit[*1,2] Stanisław Jastrzębski[*3] Nicolas Ballas[*1,2] David Krueger[*1,2] Emmanuel Bengio[4]
Maxinder S. Kanwal[5] Tegan Maharaj[1,6] Asja Fischer[7] Aaron Courville[1,2,8] Yoshua Bengio[1,2,9]
Simon Lacoste-Julien[1,2]

Generalization in machine learning

- Generalization vs. Memorization: Some directions
  - Neural networks are **content-aware** when it comes to learning.

Generalization in machine learning

- Generalization vs. Memorization: Some directions
  - Neural networks are **content-aware** when it comes to learning.
  - Randomly perturbed data and corrupted labels make the content-awareness irrelevant.

Generalization in machine learning

- Generalization vs. Memorization: Some directions
  - Neural networks are **content-aware** when it comes to learning.
  - Randomly perturbed data and corrupted labels make the content-awareness irrelevant.
    - 👆 This leads the network to memorization.

Summary so far

- Neural networks are content-aware.
- For real data, neural nets exploit patterns.
- For random stuff, neural nets tend to memorize the noise to minimize loss.

Deep Learning & Information Theory

Opening the black box of Deep Neural Networks via Information

Ravid Schwartz-Ziv RAVID.ZIV@MAIL.HUJI.AC.IL
Edmond and Lilly Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel

Naftali Tishby* TISHBY@CS.HUJI.AC.IL
School of Engineering and Computer Science
and Edmond and Lilly Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel

Mutual information: What’s that?

- Measures how much one **random variable** tells about the other.

Mutual information: What’s that?

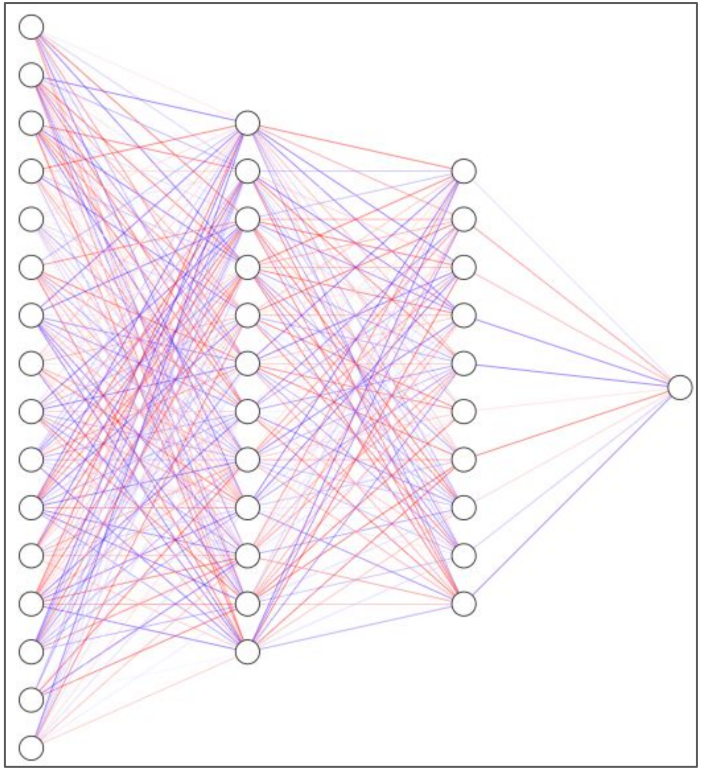
- Measures how much one **random variable** tells about the other.
- High mutual information -> Low uncertainty and vice versa.

Mutual information: What’s that?

- Measures how much one random variable tells about the other.
- High mutual information -> Low uncertainty and vice versa.
- Zero mutual information -> Variables are independent.

Variable of interest in a neural net?

- Activations



How much information is there in layer N about the input?

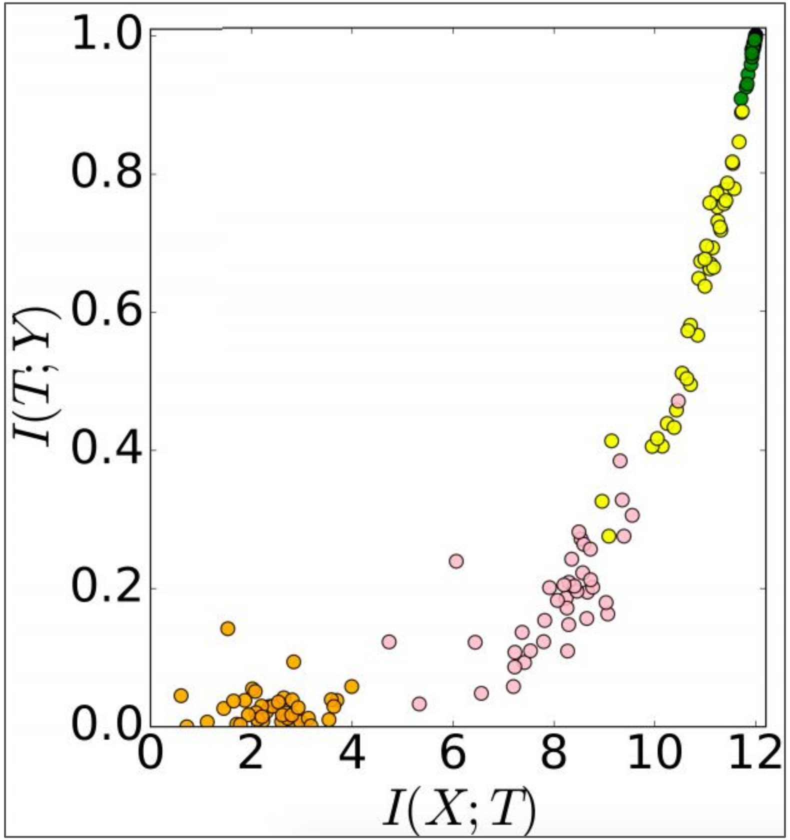
Information propagation in neural nets

- The mutual information about the input decreases successively.

Information propagation in neural nets

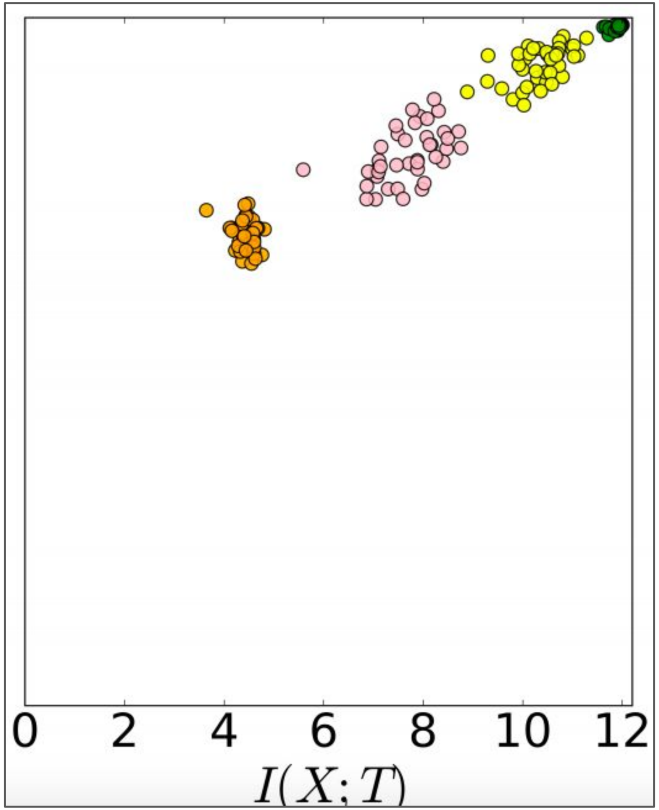
- The mutual information about the input decreases successively.
- Input contains the highest mutual information about instances and labels.

Information propagation in neural nets



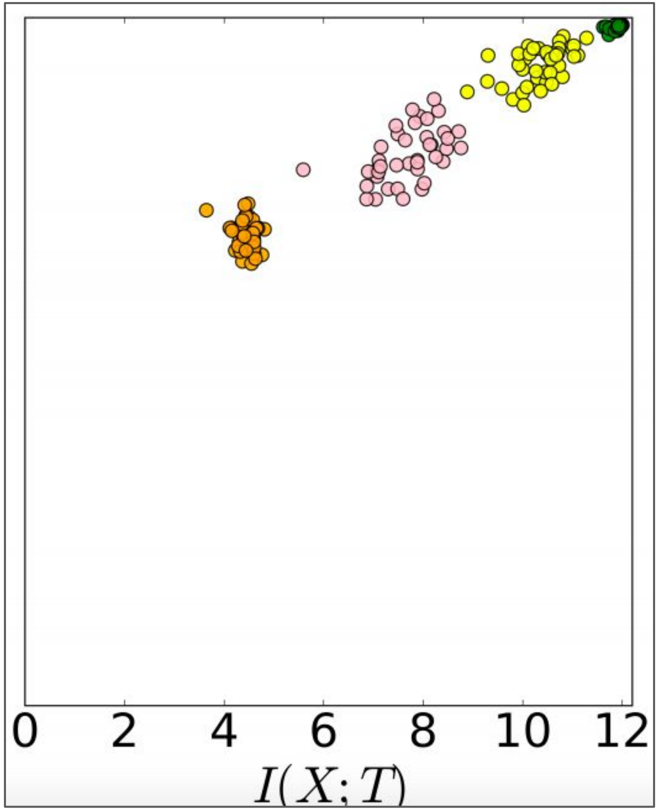
sayak.dev

Let’s start training ...



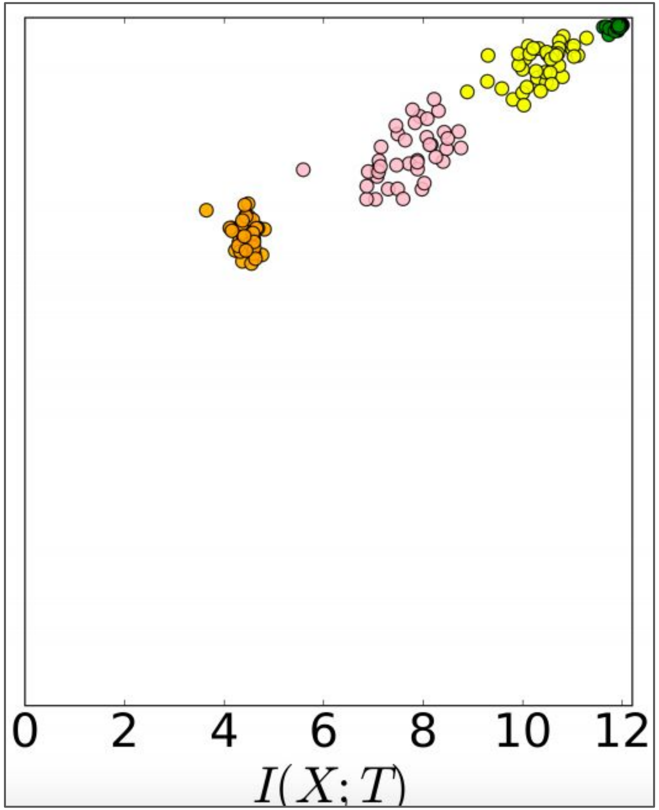
sayak.dev

Let's start training ...



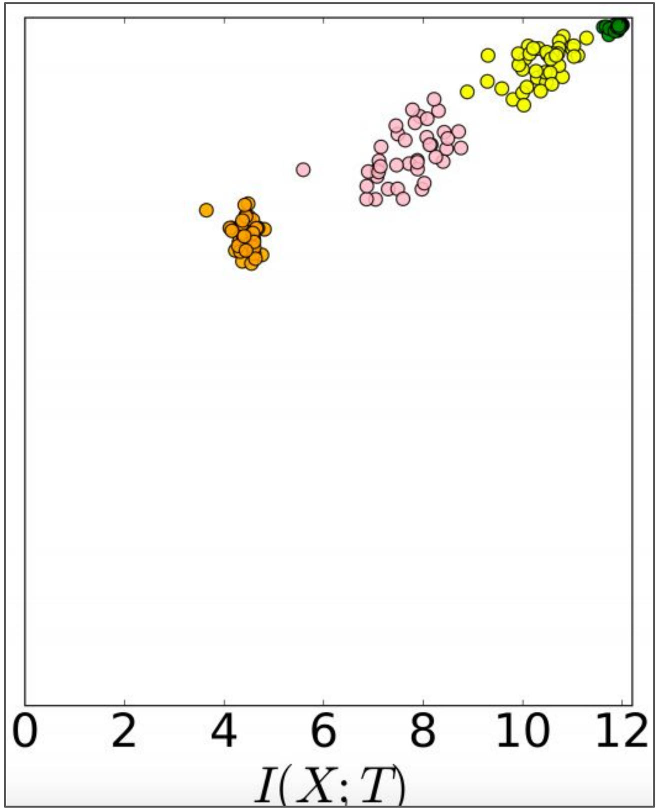
- Activations learning about the labels.

Let’s start training …



- Activations learning about the labels.
- Activations starting to **memorize** the input data.

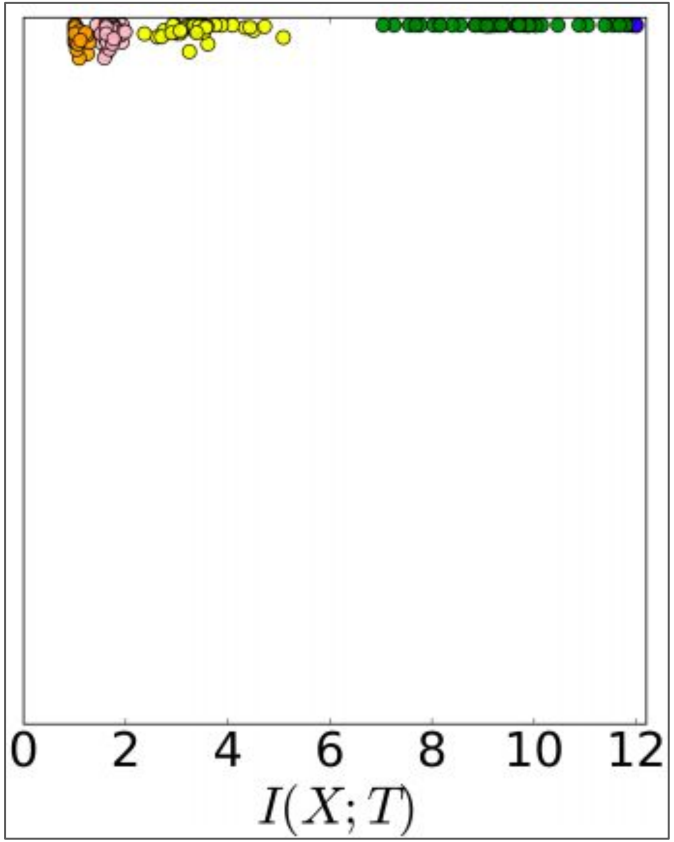
Let’s start training ...



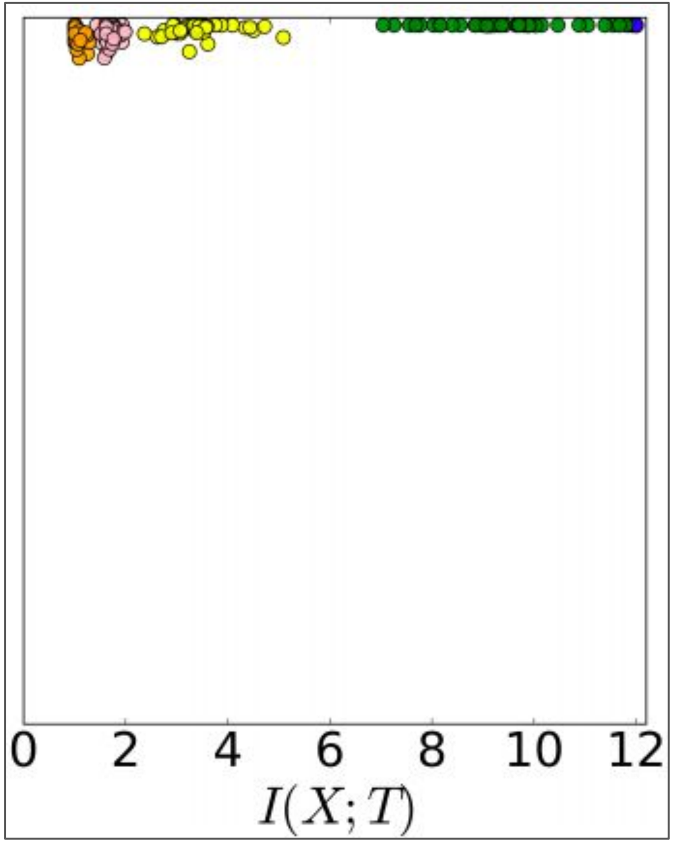
- Activations learning about the labels.
- Activations starting to **memorize** the input data.

Fitting phase!

We are still training ...

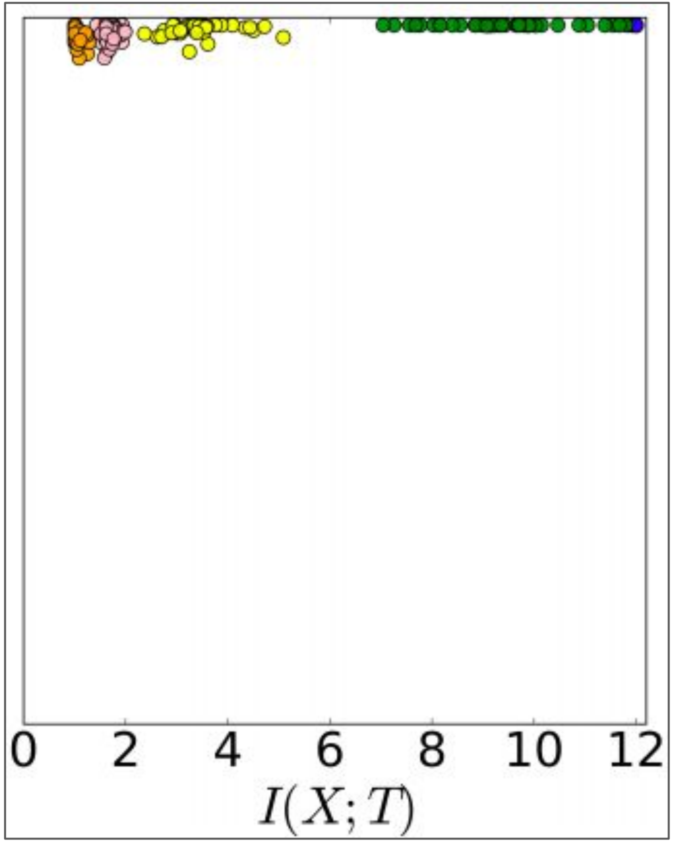


We are still training ...



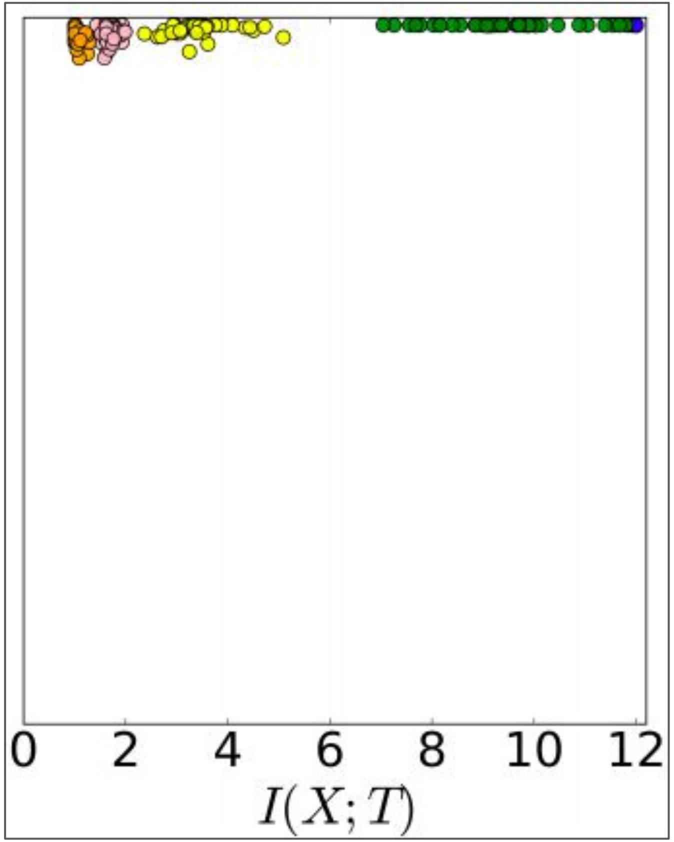
- Activations starting to discard information about input data.

We are still training ...



- Activations starting to discard information about input data.
- Activations trying to ignore the irrelevant parts of the input data.

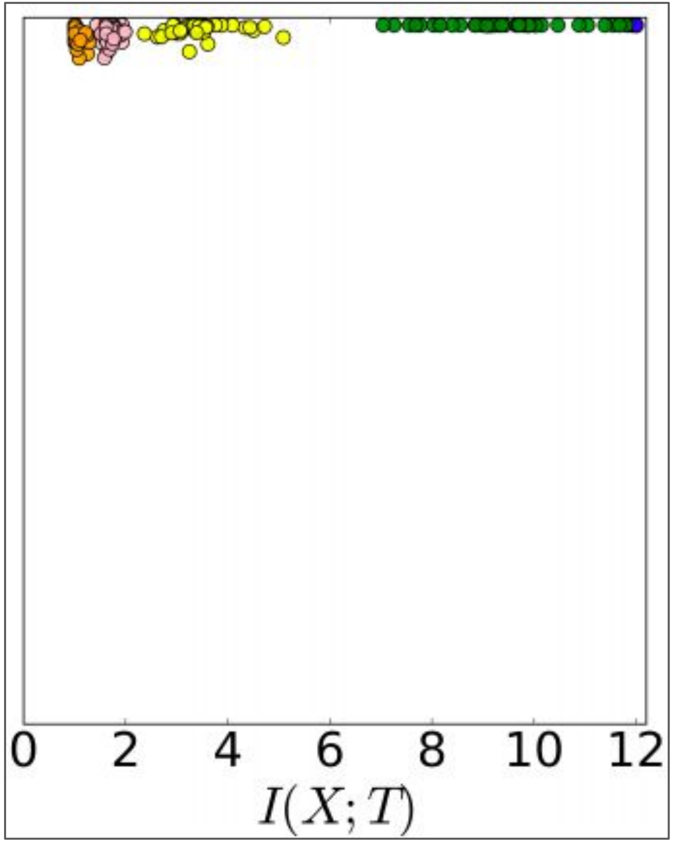
We are still training ...



- Activations starting to discard information about input data.
- Activations trying to ignore the irrelevant parts of the input data.

Forgetting phase!

We are still training ...



- Activations starting to discard information about input data.
- Activations trying to ignore the irrelevant parts of the input data.

Forgetting phase!

Forgetting phase is **slower** than fitting phase.

Information through subsets of data

- The preceding story still holds on batches of data as long as there is sufficient mutual information about data and the labels. (Larger batch size)

Information through subsets of data

- The preceding story still holds on batches of data as long as there is sufficient mutual information about data and the labels. (Larger batch size)
- For very small batches the mutual information about data and the labels tend to be less.

Information loss in the order of data size

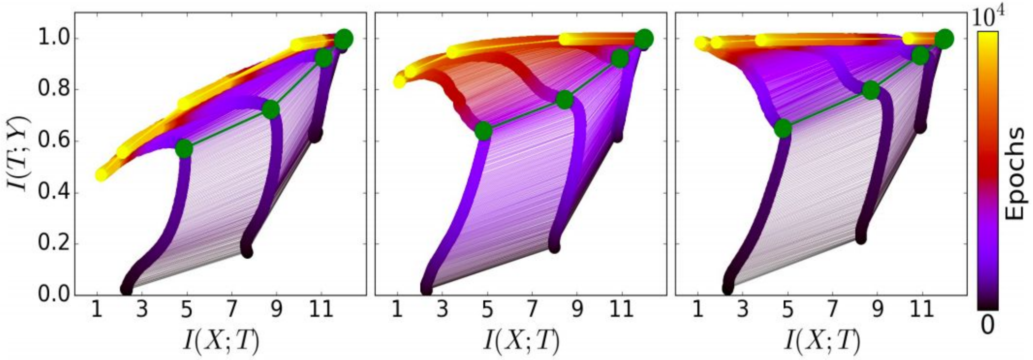
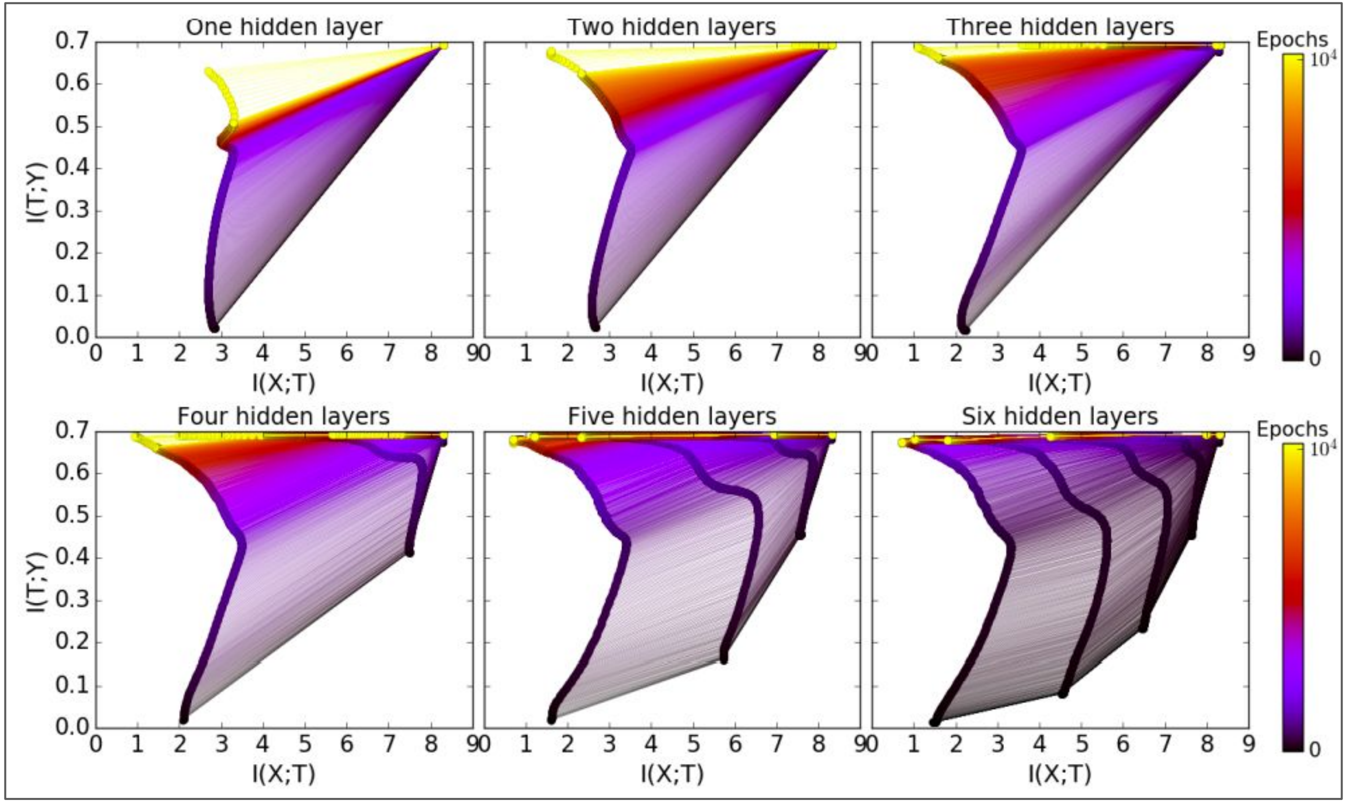


Figure 3: The evolution of the layers with the training epochs in the information plane, for different training samples. On the left - 5% of the data, middle - 45% of the data, and right - 85% of the data. The colors indicate the number of training epochs with Stochastic Gradient Descent from 0 to 10000. The network architecture was fully connected layers, with widths: input=12-10-8-6-4-2-1=output. The examples were generated by the spherical symmetric rule described in the text. The green paths correspond to the SGD drift-diffusion phase transition - grey line on Figure 4

Welcome to overfitting or **overcompression!**

The phenomenon with information loss has been referred to as Overfitting / Overcompression (by Tishbi) where we are trying to compress the data representation beyond a limit.

The beauty of depth



sayak.dev

To summarize

- Neural nets have a tendency towards memorization.
- Content awareness makes a set of input examples easier for a network to infer on.
- Information decreases as we go deeper in the network.
- With less data and bigger network the data representation gets compressed which leads to overfitting.

Some additional resources

- Toward Theoretical Understanding of Deep Learning by Sanjeev Arora
- Information Theory of Deep Learning by Naftali Tishby
- Dynamics of Neural Networks by Rajarshee Mitra

Slides available here:

http://bit.ly/kaggle-days-sayak

See you next time

Find me here:
sayak.dev

Thank you very much :)

Experts

pyimagesearch