Advanced Machine Learning

Paper: Shwartz-Ziv et al., *Opening the Black Box of Deep Neural Networks via Information* [7, 8, 9]

Gandara V. Eduardo and Werenne Aurélien

University of Liège

April 22, 2019

Overview

1. Background
2. Opening the black box
3. Discussion

Overview

1. Background

2. Opening the black box

3. Discussion

- **Information**

$$I(X) = -\log p(x)$$

Background

Information theory

- **Information**

$$I(X) = -\log p(x)$$

- **Entropy**

$$H(X) = -\sum_{x \in \mathbb{X}} p(x) \log p(x)$$

- **Information**

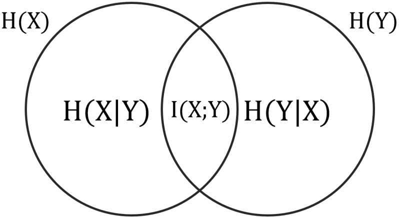
$$I(X) = -\log p(x)$$

- **Entropy**

$$H(X) = -\sum_{x \in \mathbb{X}} p(x) \log p(x)$$

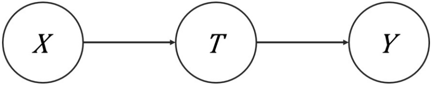
- **Mutual Information**

$$I(X;Y) = H(X) - H(X|Y)$$



- **Digital Processing Inequality:**

  For a Markov Chain of the following form,



  the inequality $I(X;T) \geq I(X;Y)$ is valid.

- **Minimal Sufficient Statistics:**

$$T(x) = \underset{S(X):\, I(S(X);Y) = I(X;Y)}{\arg\min} I(S(X);X)$$

which can be rewritten with as the following Lagrangian,

$$\min_{p(x),p(y|t),p(t)} I(X;T) - \beta I(T;Y)$$

Overview

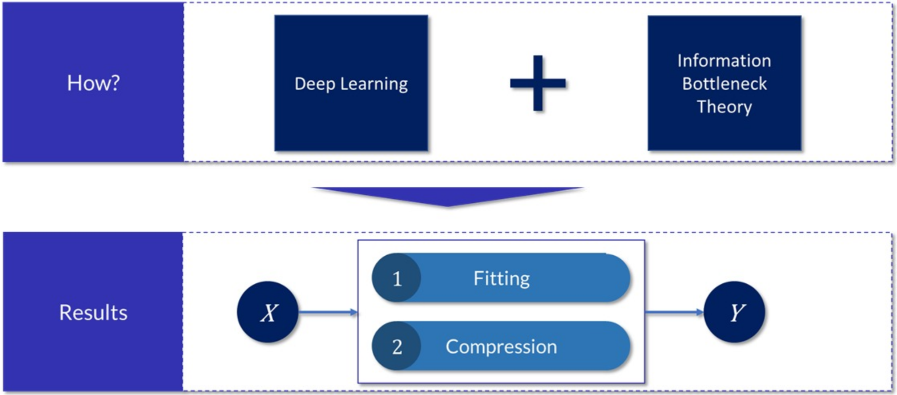
1 Background

2 Opening the black box

3 Discussion

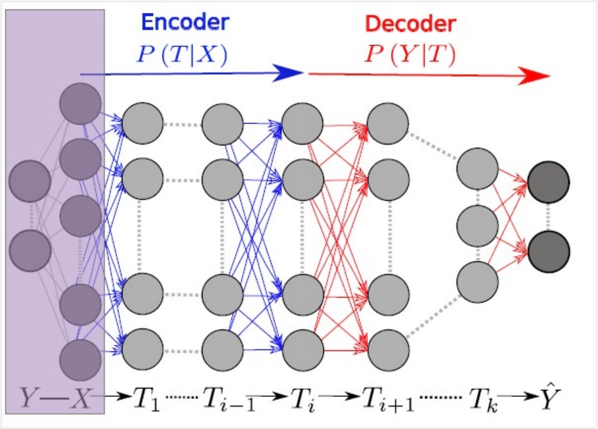
Opening the black box

Main idea



Opening the black box

Framework

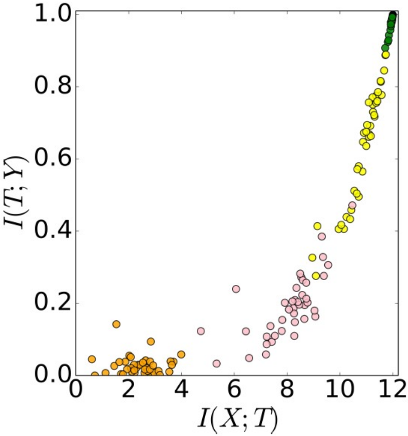

Encoder
$P(T|X)$
Decoder
$P(Y|T)$
$Y - X \rightarrow T_1 \cdots T_{i-1} \rightarrow T_i \rightarrow T_{i+1} \cdots T_k \rightarrow \hat{Y}$

$$\text{Supervised Deep Learning} \overset{?}{=} \min\{I(X;T) - \beta I(T;Y)\}$$

Opening the black box

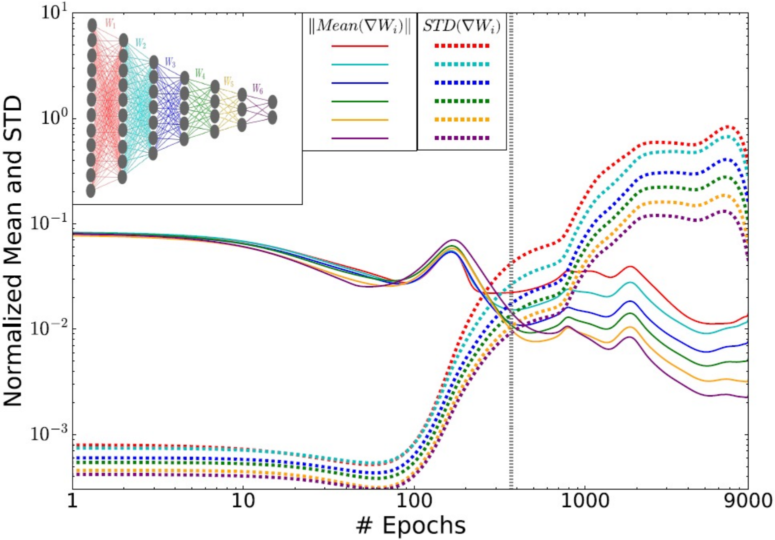
Information Plane

- Information plane (video)



Opening the black box

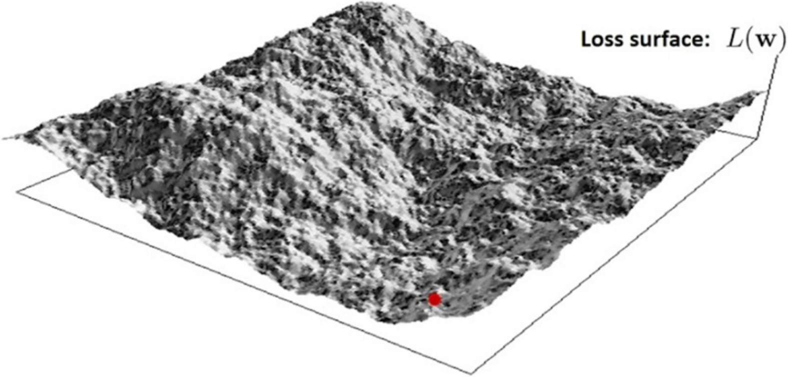
Norm and Variance of Gradients



Opening the black box

Flat Minima

Continuing SGD in local minima can be modelled as a Random Walk!



Opening the black box

Diffusion

Why can a Random Walk be seen as a compression?

Opening the black box

Diffusion

Why can a Random Walk be seen as a compression?

$\Rightarrow$ Increases entropy of weights

Opening the black box

Diffusion

Why can a Random Walk be seen as a compression?

⇒ Increases entropy of weights

⇒ Maximizes $H(X|T_i)$

Opening the black box
Diffusion

Why can a Random Walk be seen as a compression?

$\Rightarrow$ Increases entropy of weights

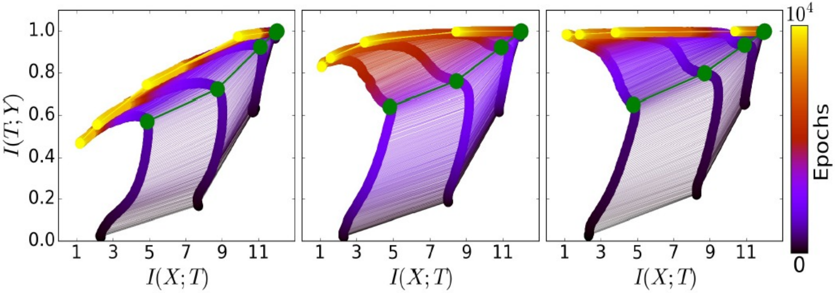
$\Rightarrow$ Maximizes $H(X|T_i)$

$\Rightarrow$ Minimizes $I(X;T_i) = \underbrace{H(X)}_{constant} - H(X|T_i)$

Opening the black box

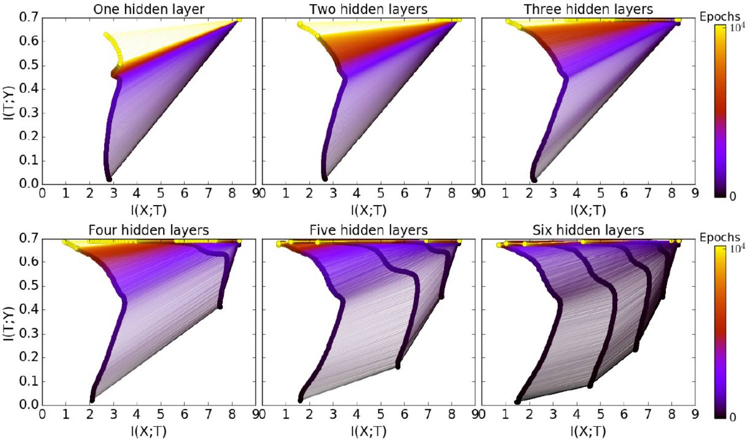
Underfitting & overfitting

- Evolution in the information plane, for different training sample sizes. From left to right (5% of the data, 45% of the data, 85% of the data).



Opening the black box

Hidden Layers

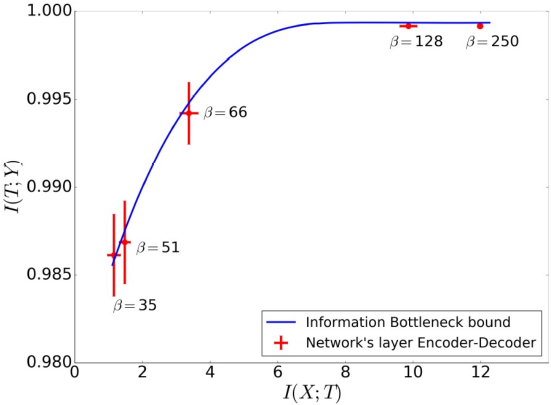


Why? $\Rightarrow$ $\boxed{exp(\sum_k \Delta I_X^k) \gg \sum_k exp(\Delta I_X^k)}$

Opening the black box

IB Curve

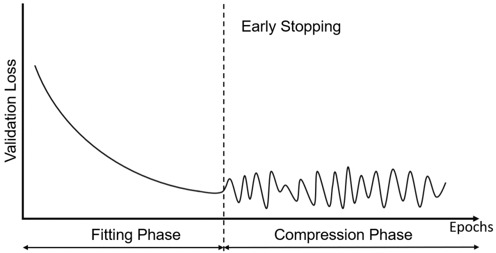
$$\text{Supervised Deep Learning} \approx \min\{I(X;T) - \beta I(T;Y)\}$$



Opening the black box

Summary

- Fitting then Compressing
- Overfitting = Overcompression
- Noise of SGD ⇒ Generalization
- Do we need to adapt *Early Stopping* method?



Overview

1 Background

2 Opening the black box

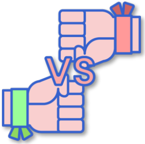
3 Discussion

Discussion

Controversy

Advocates

- New MI estimators [3,6]
- Relationship with VAE [2]
- Train longer, Generalize better [4]



Detractors

- Saxe et al. [5]
- Hard to reproduce results

Discussion

Pros and cons

Pros and cons:

- ✓ New insight of how DL works
- ✓ Led to various applications [1, 2]
- ✓ Justified Mathematically
- ✓ Available code

- × Estimating MI is computationally expensive
- × Not tested on non-saturating functions and large networks
- × No verification of the results on challenging datasets

Alessandro Achille and Stefano Soatto.
Information Dropout: Learning Optimal Representations Through Noisy Computation.
arXiv e-prints, page arXiv:1611.01353, Nov 2016.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy.
Deep Variational Information Bottleneck.
arXiv e-prints, page arXiv:1612.00410, Dec 2016.

Ivan Chelombiev, Conor Houghton, and Cian O'Donnell.
Adaptive Estimators Show Information Compression in Deep Neural Networks.
arXiv e-prints, page arXiv:1902.09037, Feb 2019.

Elad Hoffer, Itay Hubara, and Daniel Soudry.
Train longer, generalize better: closing the generalization gap in large batch training of neural networks.
arXiv e-prints, page arXiv:1705.08741, May 2017.

Andrew et al. M. Saxe.
On the information bottleneck theory of deep learning.
ICLR 2018, 2015.

Morteza Noshad, Yu Zeng, and III Hero, Alfred O.
Scalable Mutual Information Estimation using Dependence Graphs.
arXiv e-prints, page arXiv:1801.09125, Jan 2018.

Ravid Shwartz-Ziv and Naftali Tishby.
Opening the Black Box of Deep Neural Networks via Information.
arXiv e-prints, page arXiv:1703.00810, Mar 2017.

Naftali Tishby, Fernando C. Pereira, and William Bialek.
The information bottleneck method.
arXiv e-prints, page physics/0004057, Apr 2000.

Naftali Tishby and Noga Zaslavsky.
Deep Learning and the Information Bottleneck Principle.
arXiv e-prints, page arXiv:1503.02406, Mar 2015.