Information Bottleneck (and Unsupervised Learning)

Fabio Massimo Zennaro
fabiomz@ifi.uio.no

University of Oslo

Information Bottleneck

Information bottleneck [14] is a *information-theoretic* framework for learning.

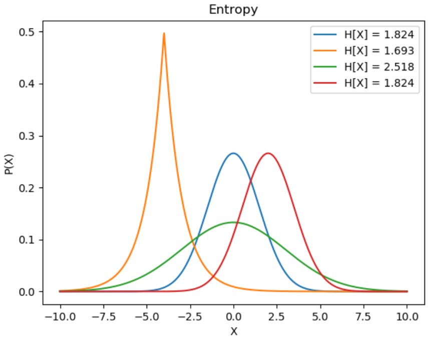
- Simple and elegant
- It can be used to *explain* learning [10]
- It can be used to *direct* learning [1]
- It is computationally non-trivial [3, 2, 8]

Entropy

Entropy of a random variable X:

$$H[X] = -\sum_x p(x) \log p(x)$$

- Statistical descriptor
- Domain-insensitive
- Measure of information
- Measure of uncertainty
- Measure of concentration



Mutual Information

Mutual information of two random variables X, Y:

$$
\begin{aligned}
I[X;Y] &= H[X] - H[X|Y] \\
&= H[Y] - H[Y|X]
\end{aligned}
$$

- Invariant to invertible reparametrization
- Measure of shared information
- Measure of reduction of uncertainty

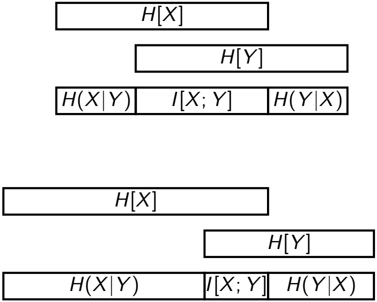


Diagram from [5]

The Learning Problem (1)

We phrase the *learning problem* as a *mapping* problem:

$$X \to Y$$

- X,Y are two (potentially high-dimensional) variables
- X may be images/genomes/videoframes,
  Y may be categorical labels/expression levels/reward signals.

The Learning Problem (2)

Further, let us assume that we may solve the *learning problem* using *intermediate representations*:

$$X \to Z \to Y$$

- Intermediate representation inspired by real(!) neural networks
- Z encodes efficiently X (*compression*)
  Z eases mapping onto Y (*relevance*)

The Information Bottleneck (1)

How can we compute *optimal intermediate representations* Z?

We want to Z to contain all and only the information relevant to Y:

$$\min \underbrace{I[X;Z]}_{\text{compression}} \qquad \max \underbrace{I[Z;Y]}_{\text{relevance}}$$

- We maximize the compression by *minimizing the mutual information between X and Z*
- We maximize the relevance by *maximizing the mutual information between Z and Y* (Infomax principle [4])
- (Information theory $\rightarrow$ *rate-distortion theory*)
- (Statistics $\rightarrow$ *sufficient statistics*)

The Information Bottleneck (2)

We can re-express our objective as a *single optimization problem*:

$$\arg\min_{Z} I[X;Z] - \beta I[Z;Y]$$

- Optimization is wrt Z (it may be a parametric representation)
- β is a Lagrangian and trades off compression and relevance
- (This has an analytic solution using *Blahut-Arimoto algorithm* [14])
- (Practically, though, estimating mutual information is hard)

Applications

Two main ways of using the IB:

1. Analyzing existing algorithms [10, 7]
2. Plugging it in existing algorithms [1]

We will focus on the first one.

Opening the Black Box of DNN via IB (1)

Can we explain learning in deep neural networks using IB? [10]

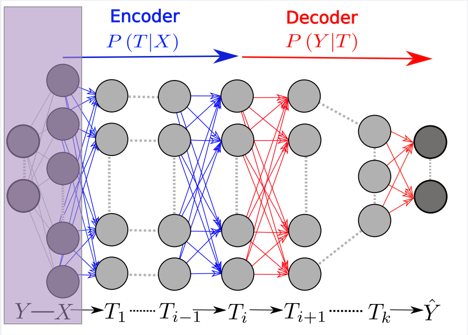


Image from [10]

- Every layer of the network computes an intermediate representation Z_i

Opening the Black Box of DNN via IB (2)

Can we explain learning in deep neural networks using IB? [10]

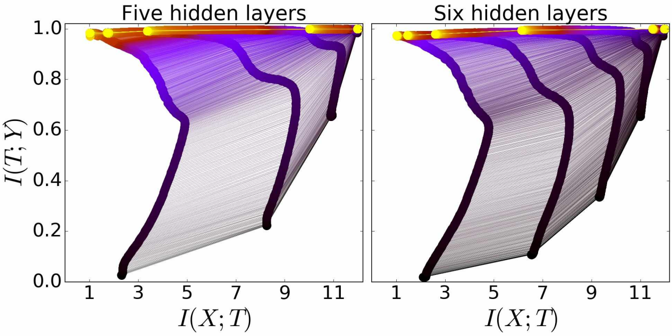


Image from [10]

- Trajectory in the information plane agrees with IB theory
- (Two different learning phases may be identified)
- (There are some criticisms of this analysis [9])

IB and Unsupervised Learning

Can we use IB theory to analyze UL algorithms?

- UL algorithms do not have specific target Y
- How do we define *relevance*?
- Need other measures/constraints

IB and Sparse Filtering (1)

Let us take **sparse filtering**, a UL algorithm to learn *maximally sparse representation* of the data [6].

- Sparsity has a strong biological inspiration
- Not totally clear why it works [15]

It has been suggested that sparse filtering solves the following *information theoretic problem* [15]:

$$\arg\max_{Z} I\left[X;Z\right] + H\left[Z\right]$$

IB and Sparse Filtering (2)

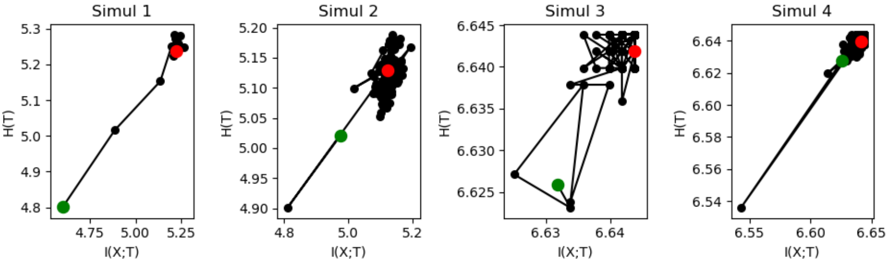


Image from [16]

- Preliminary results seem to agree with the hypothesis
- How to *connect* with IB theory?
- How *generalizable* to SF and UL in general are these results?
- What *insights* on SF and UL can we get?

Conclusions

- IB is a very general theory of learning
- There are alternative information bottleneck formulations [11, 13]
- This is not the only information-theoretic principle we can use for learning [12]
- Application to UL may be very interesting!

Thanks

Thank you for listening!

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[3] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[4] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[5] David J.C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.

[6] Jiquan Ngiam, Zhenghao Chen, Sonia A Bhaskar, Pang W Koh, and Andrew Y Ng. Sparse filtering. In *Advances in neural information processing systems*, pages 1125–1133, 2011.

[7] Nikolaos Nikolaou, Henry Reeve, and Gavin Brown. Margin maximization as lossless maximal compression. *arXiv preprint arXiv:2001.10318*, 2020.

[8] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

[9] Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. On the information bottleneck theory of deep learning. 2018.

[10] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

References III

[11] Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. In *Advances in neural information processing systems*, pages 617–623, 2000.

[12] Greg Ver Steeg. Unsupervised learning via total correlation explanation. *arXiv preprint arXiv:1706.08984*, 2017.

[13] DJ Strouse and David J Schwab. The deterministic information bottleneck. *Neural computation*, 29(6):1611–1630, 2017.

[14] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[15] Fabio Massimo Zennaro and Ke Chen. Towards understanding sparse filtering: A theoretical perspective. *Neural Networks*, 98:154–177, 2018.

References IV

[16] Fabio Massimo Zennaro and Ke Chen. Towards further understanding of sparse filtering via information bottleneck. *arXiv preprint arXiv:1910.08964*, 2019.