Deep Learning and the Information Bottleneck Principle
— The Informaiton Bottleneck Method

Presenter: Baosong Yang

Basic Questions:

- The design principles of deep networks are not well understood
  - The optimal achitecture
  - The number of reuqired layers
  - The sample complexity
  - The best optimiztion algorithms

- Methods: Informaiton Battleneck Principle
  - *A new idea called "information bottleneck" is helping to explain the puzzling success of today's AI algorithms, and might also explain how human brains learns.* — 《New Theory Cracks Open the Black Box of Deep Learning》

Information Bottleneck

- The goal of informaiton bottleneck:
  - Maximally compresse input and preseves as much as possible the information on output. $Y \to X \to \hat{X}$
  - Minimize the mutual information $I\left(X;\hat{X}\right)$ to obtain the simplest statistics under a constraint on $I\left(\hat{X};Y\right)$

$$\mathcal{L}\left[p\left(\hat{x}|x\right)\right] = I\left(X;\hat{X}\right) - \beta I\left(\hat{X};Y\right)$$

  - Tradeoff between the complexity of compressed input and perseved relevant information.
  - $\hat{X}$ not exist => varitional problem

- The IB variational problem satisfy the following self-consistent equations, which can be iterated:

$$\begin{aligned} p(\hat{x}|x) &= \frac{p(\hat{x})}{Z(x;\beta)} \exp\left(-\beta D\left[p(y|x) \| p(y|\hat{x})\right]\right) \\ p(y|\hat{x}) &= \sum_x p(y|x)\, p(x|\hat{x}) \\ p(\hat{x}) &= \sum_x p(x)\, p(\hat{x}|x) \end{aligned}$$

- The residual information between X and Y (the relevant information not captured by $\hat{X}$)

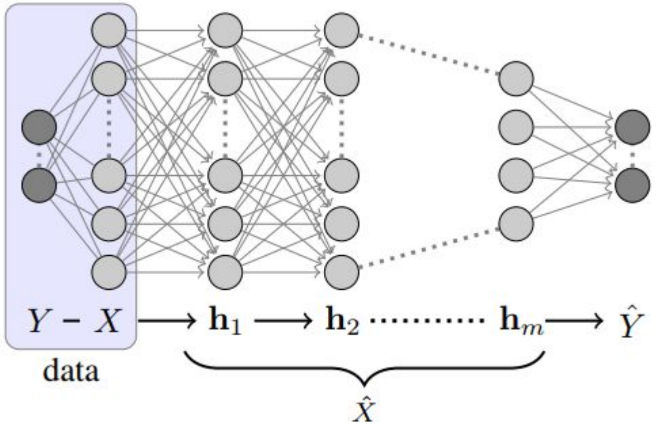
$$D_{IB} = E\left[d_{IB}\left(X, \hat{X}\right)\right] = I(X; Y|\hat{X})$$

- The variational principle is equivalent to:

$$\tilde{\mathcal{L}}\left[p(\hat{x}|x)\right] = I\left(X; \hat{X}\right) + \beta I\left(X; Y|\hat{X}\right)$$

DNNs

- The goal of supervised learning:
  - Extracts an approximate minimal sufficient statistics of the input with respect to the output.



- Most of the entropy of X is not very informative about Y
- DNNs sequentially process X to Y (Markov chain)

Information characteristics of the layers

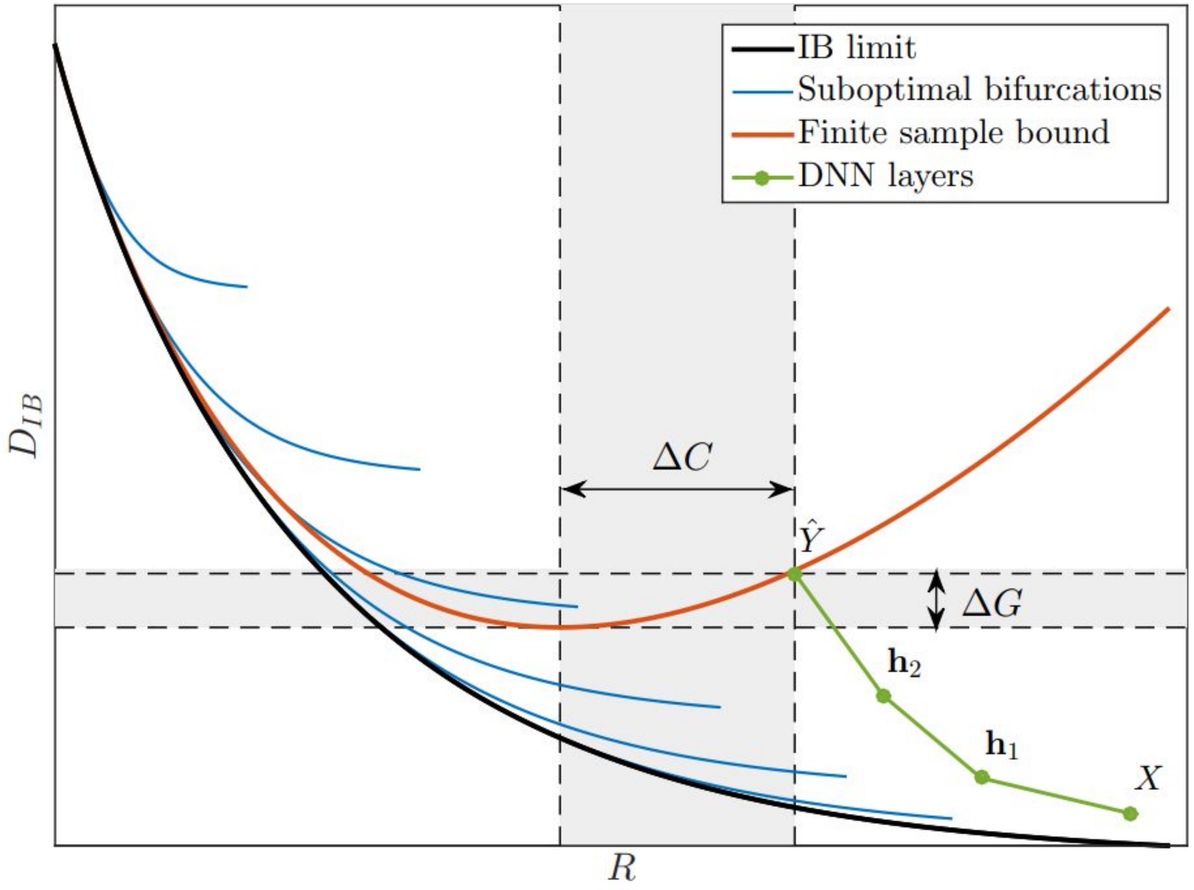
- Data Processing Inequality: Information about Y that is lost in one layer connot be recovered in higher layers.

$$I(Y;X) \geq I(Y;\mathbf{h}_j) \geq I(Y;\mathbf{h}_i) \geq I\left(Y;\hat{Y}\right)$$

- Each layer should attempt to maximize $I(Y;\mathbf{h}_i)$ while minimizing $I(\mathbf{h}_{i-1};\mathbf{h}_i)$

- $\tilde{\mathcal{L}}[p(\hat{x}|x)] = I\left(X;\hat{X}\right) + \beta I(X;Y|\hat{X})$ => $I(\mathbf{h}_{i-1};\mathbf{h}_i) + \beta I(Y;\mathbf{h}_{i-1}|\mathbf{h}_i)$

Finite Samples and Generalization Bounds



- Finite Samples

$$I\left(\hat{X};Y\right) \leq \hat{I}\left(\hat{X};Y\right) + O\left(\frac{K\,|\mathcal{Y}|}{\sqrt{n}}\right)$$

$$I\left(X;\hat{X}\right) \leq \hat{I}\left(X;\hat{X}\right) + O\left(\frac{K}{\sqrt{n}}\right)$$

- Generalization Bounds

$$K \approx 2^{I(\hat{X};X)}$$

- Generalization gap: the amount of information about Y did not capture

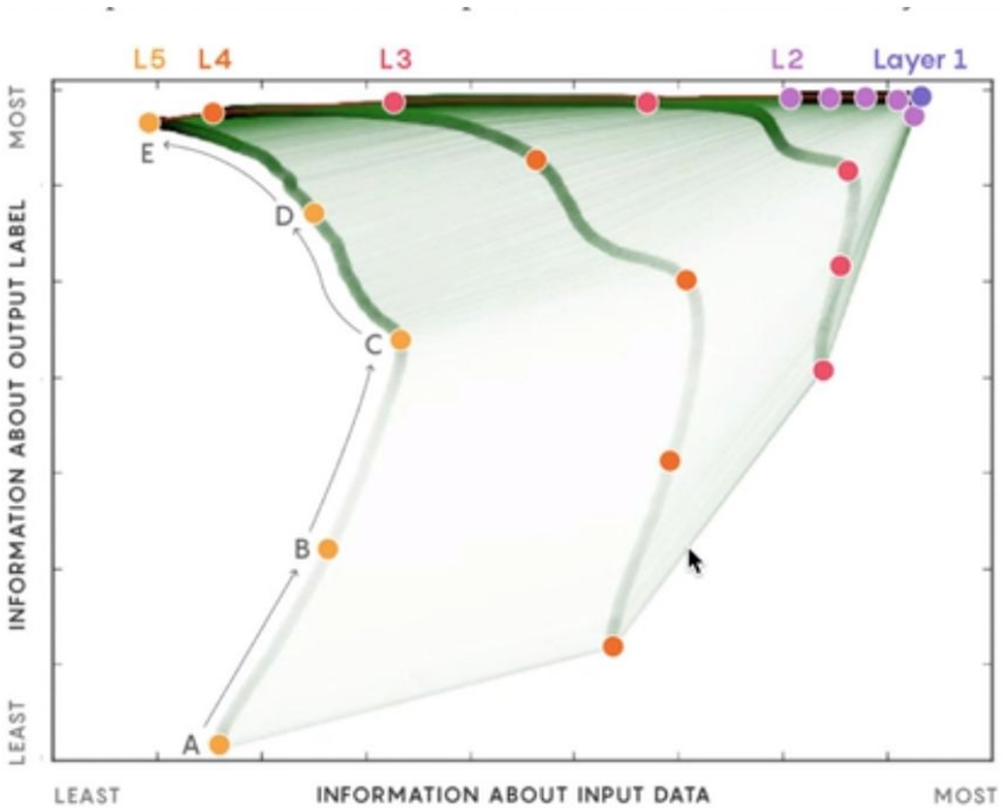
$$\Delta G = D_N - D_{IB}^{*}(n)$$

- Complexity gap: the amount of unnecessary complexity in the network

$$\Delta C = R_N - R^{*}(n)$$

Proof at: Learning and generalization with the information bottleneck

The principle of DNNs: Forget and Preserve



Opening the Black Box of DNNs via Informaiton, arXiv 2017

Recommended Readings

- Tishby's talk about Information Bottleneck for DNNs
  - Video: https://www.youtube.com/watch?v=bLqJHjXihK8&t=262s
  - 解析：https://blog.csdn.net/qq_20936739/article/details/82453558
- Deep Learning and the information Bottleneck Principle
  - 中文解析（Video）:https://www.bilibili.com/video/av18080637
- New Theory Cracks Open the Black Box of DL
  - https://www.quantamagazine.org/new-theory-cracks-open-the-black-box-of-deep-learning-20170921/
- 论文精读
  - https://blog.csdn.net/qq_25011449/article/details/81258919