On the Information Bottleneck Theory of Deep Learning

Attempts on Opening the Black Box of Deep Learning

Presented by Julaiti Alafate

Nov 20, 2018

Plan

Background

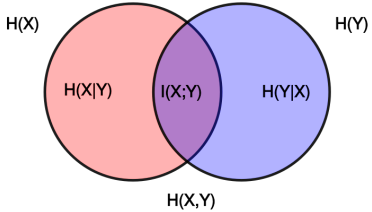
Opening the Black Box of DNN

Different Opinion

Conclusion

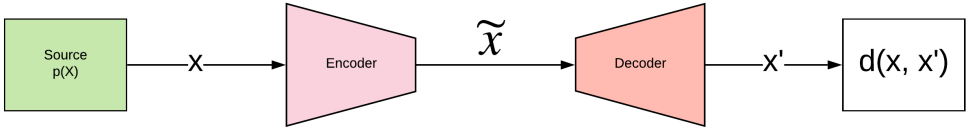
Informal Overview of the Information Theory

- Entropy $H(X)$: the average number of bits to describe the outcome of a random variable X
- Conditional Entropy $H(Y|X)$: the average number of bits to describe the outcome of Y *given* that the value of X is known.
- Mutual Information $I(X;Y)$: the average number of bits that X contains about Y.

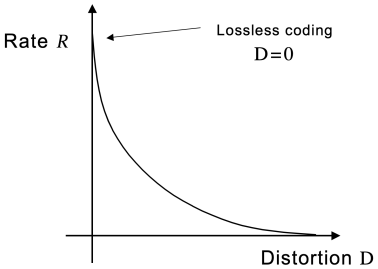


$$I(X;Y) = H(Y) - H(Y|X) \qquad (1)$$

Rate-Distortion (RD) Theory



Lower the bit-rate R by allowing some distortion D of the original "information"



$$
\begin{aligned}
R(D^*) &= \min_{D \le D^*} I(X;Y) \\
&= \min_{D \le D^*} H(X) - H(X|Y)
\end{aligned}
\tag{2}
$$

The Blahut-Arimoto Algorithm

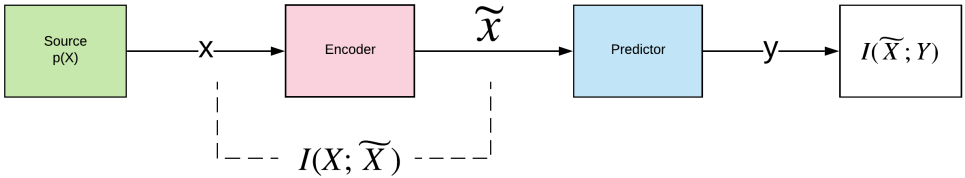
- Problem

$$\min_{p(\widetilde{x}|x)} \underbrace{I(X;\widetilde{X})}_{\text{Rate}} + \beta \underbrace{E[d(X,\widetilde{X})]}_{\text{Distortion}} \tag{3}$$

- Example: Image compression
- Solution:
  Blahut-Arimoto Algorithm - an alternating iterative algorithm to calculate $p(\widetilde{x})$ and $p(\widetilde{x}|x)$.

The Information Bottleneck (IB) Method

IB extends RD theory to prediction, by measuring the quality of the encoding by its ability to predict another random variable.



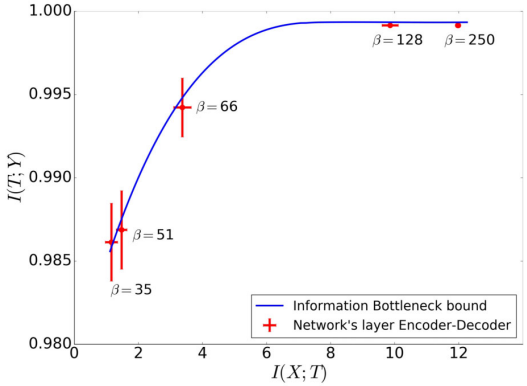
- Problem

$$\min_{p(\widetilde{x}|x)} \underbrace{I(X;\widetilde{X})}_{\text{Compression}} - \beta \underbrace{I(\widetilde{X};Y)}_{\text{Prediction}} \tag{4}$$

- Solution: Similar iterative algorithm as Blahut-Arimoto to calculate $p(\widetilde{x}|x)$, $p(\widetilde{x})$, and $p(y|\widetilde{x})$.

Information Plane

The solution to IB defines a concave curve of the optimal representation that separates the *achievable* and *unachievable* regions.



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

Plan

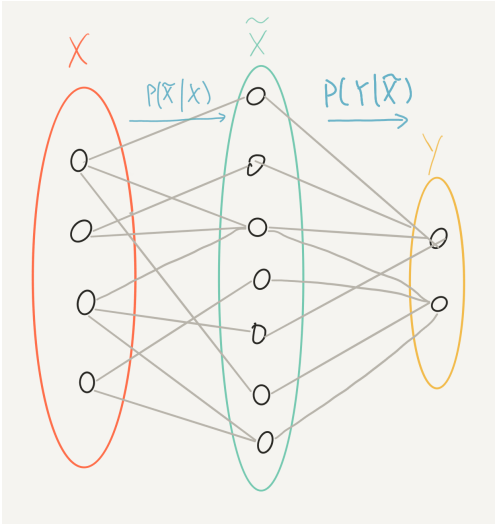
Background

Opening the Black Box of DNN

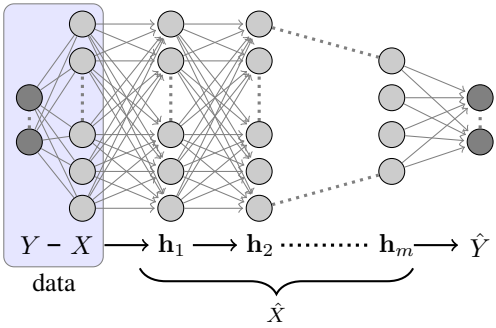
Different Opinion

Conclusion

Setup: Take the Whole Layer as a Single Random Variable



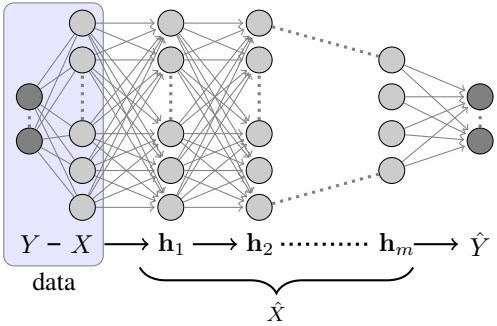
From the Information Theory Perspective



- Network layers form a Markov chain
- From data processing inequality (DPI), we know the information about Y that is lost in one layer cannot be recovered in later layers.

$$I(Y;X) \geq I(Y;h_1) \geq I(Y;h_2) \geq \ldots \geq I(Y;\widetilde{Y}) \tag{5}$$

From the Information Bottleneck (IB) Perspective



- Each layer can be compared to the optimal IB limit for some β

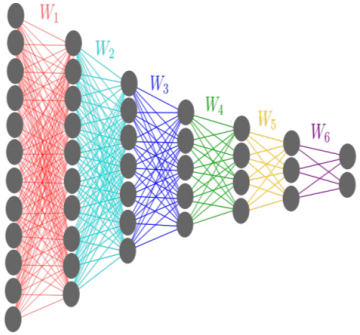
$$I(h_{i-1}; h_i) + \beta I(Y; h_{i-1}|h_i) \tag{6}$$

- Each consecutive layer can only increase the IB distortion level, and also compresses its inputs by eliminating irrelevant information

Main Results

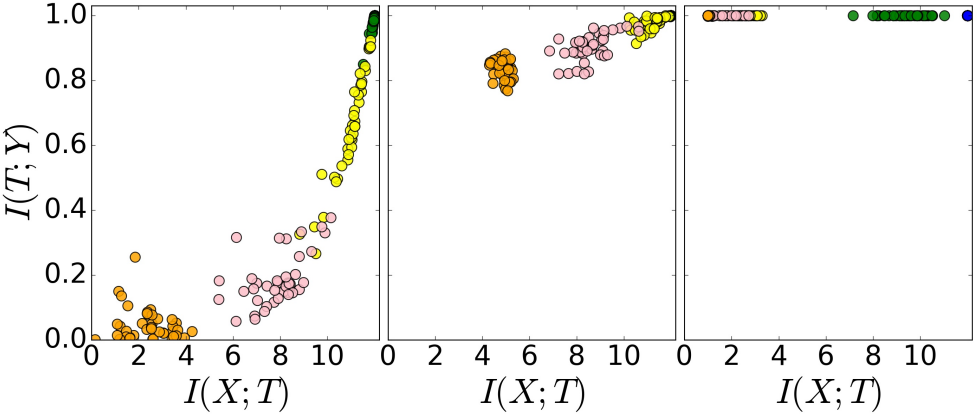
1. SGD has two main phases: first increase the information on the labels (fitting), second reduce the information on the input (compressing).
2. The main advantage of the hidden layers is reducing the compression times.
3. The converged layers lie on or very close to the IB theoretical bound.

Numerical Experiments



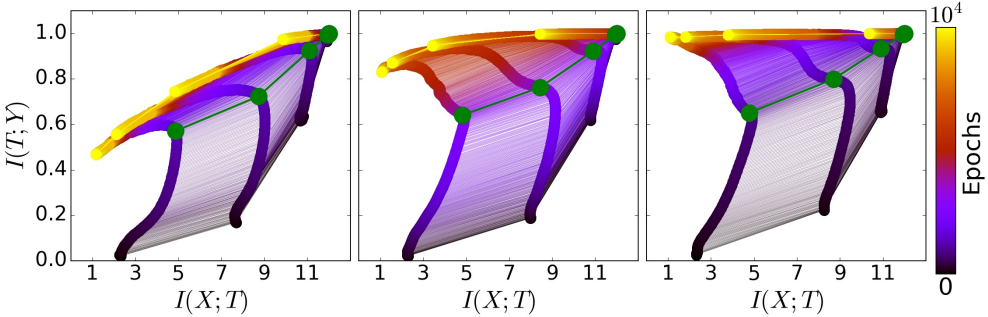
- Synthetic dataset, 12 binary inputs, 1 binary output (not deterministic)
- $P(X, Y)$ is known, $P(T|X)$ and $P(Y|T)$ can be calculated directly
- Network structure: 7 fully-connected layers, widths 12-10-7-5-4-3-2 neurons
- All neurons use $\tanh$ as the activation function
- Estimate MI: split output of the layers into 30 equal intervals in $[-1, 1]$

SGD has Two Phases



- Fast empirical risk minimization (EMR) phase
- . . . followed by the much longer representation compression phase
- Click this for an animation video

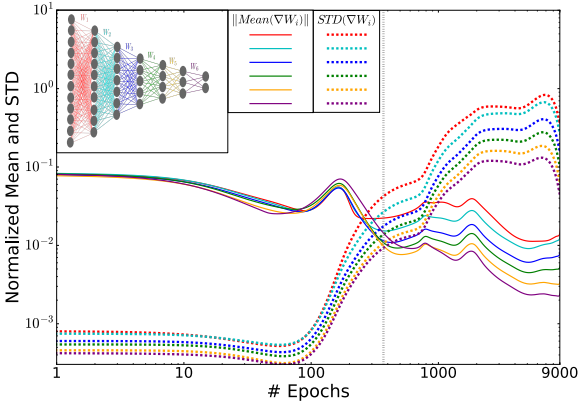
Different Sample Sizes



Training with 5%, 45%, 85% of the data, respectively

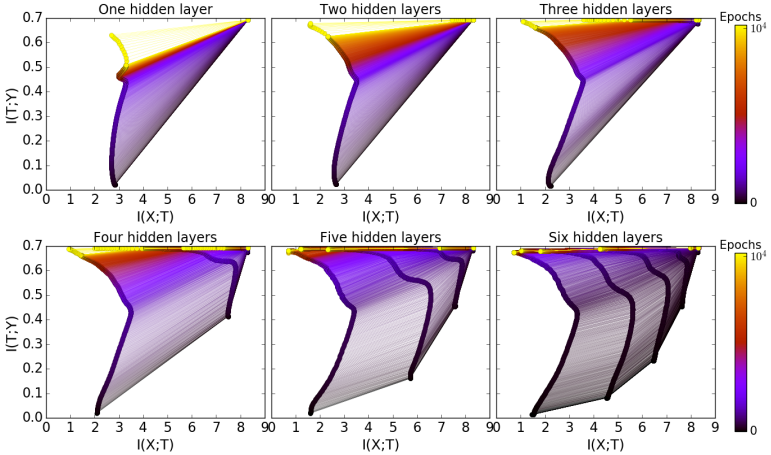
- Small sample sizes loss label information during the compression phase
- . . . causes over-fitting, which can be prevented by early stopping methods

Drift and Diffusion Phases of SGD Optimization



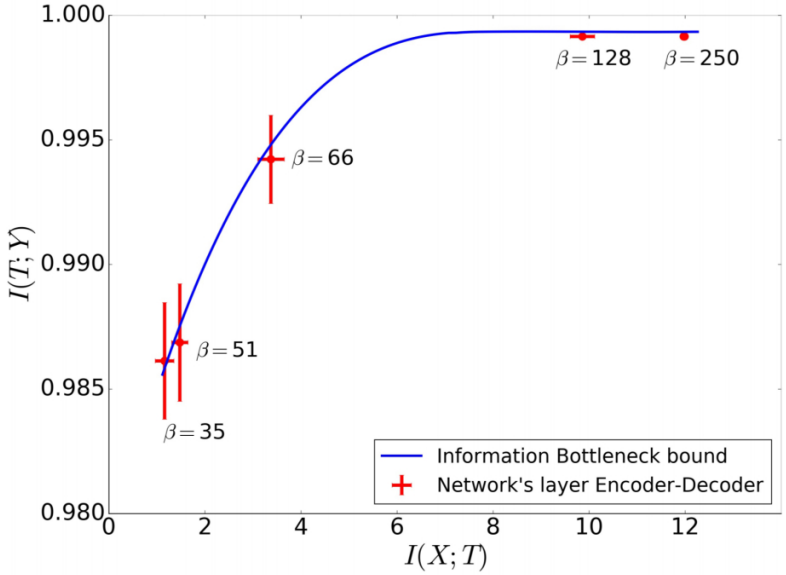
- In diffusion phase, SGD adds noise to the weights, which maximizes $H(X|T_i)$, thus minimizes $I(X;T_i) = H(X) - H(X|T_i)$.

The computational benefit of the hidden layers



- More intermediate representations accelerates the compression

Converged Layers are close to the IB Bound



Black box is opened. Problem solved?

On the Information Bottleneck Theory of Deep Learning (ICLR 2018):

Here we show that none of these claims hold true in the general case.

Plan

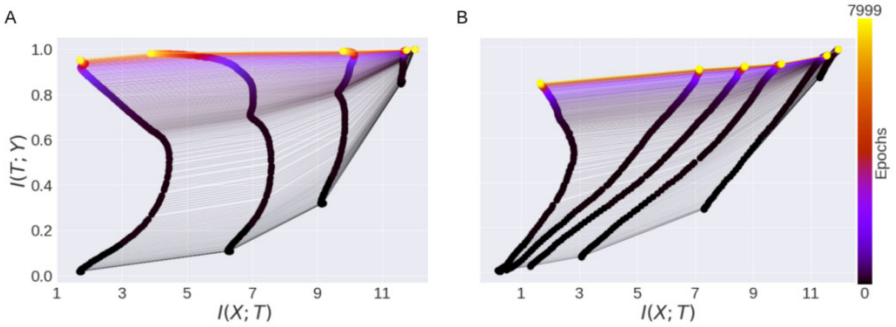
Background

Opening the Black Box of DNN

Different Opinion

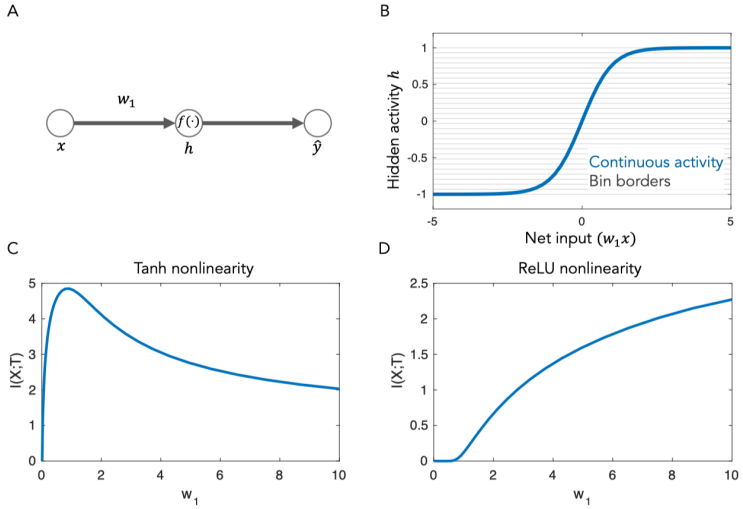
Conclusion

Compression depends on the Activation Function Choice

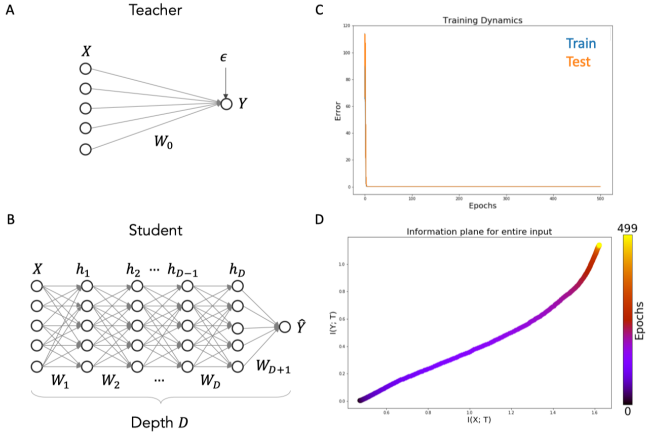


Training with $\tanh$ (left) and ReLU (right) as the activation functions

The Impact of Different Nonlinearity on the MI Dynamics

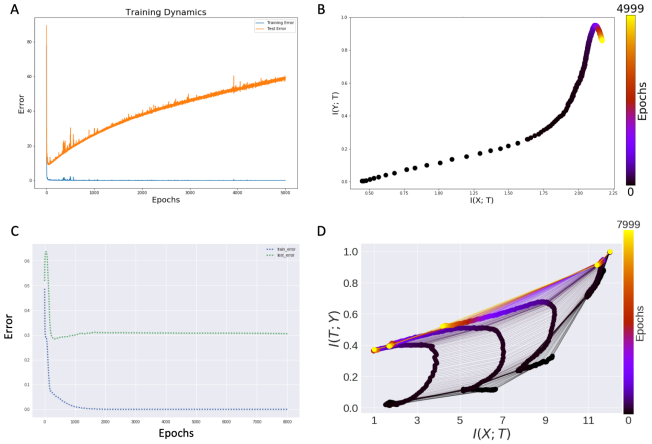


Generalisation Independent of Compression



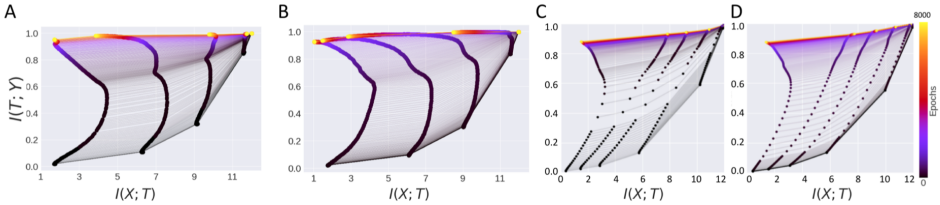
- Networks that do not compress may also generalize well.

Generalisation Independent of Compression (cont.)



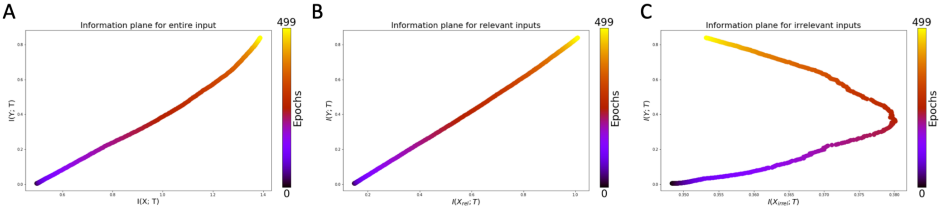
- Networks that do compress may not generalize well.

Compression Dynamics may not differ in Batch GD and SGD



$\tanh$ with SGD (A) and Batch GD (B), ReLU with SGD (C) and Batch GD (D)

Task-Irrelevant Information is Compressed



- The information with irrelevant inputs may indeed reduce, while overall MI with the input may still increase.

Plan

Background

Opening the Black Box of DNN

Different Opinion

Conclusion

Conclusion

1. Compression dynamics in the information may be a general feature of deep networks
   - or may not, but influenced by the nonlinearities employed by the network
2. Stochasticity in the training process may contribute to compression
   - or may not, because batch gradient descent shows similar compression dynamics
3. Generalization performance may not relate to the information plane behavior
   - but compression may still occur to a subset of features if the task demands it
4. Bottom line: the IB principle itself is more general and may yet offer important insights into DNN.

References

1. Tishby, Naftali, Fernando C. Pereira, and William Bialek. "**The Information Bottleneck Method**." Proceedings of the 37th Allerton Conference on Communication, Control and Computation, 2001.
2. Shwartz-Ziv, Ravid, and Naftali Tishby. "**Opening the Black Box of Deep Neural Networks via Information**," March 2, 2017.
3. Tishby, N., and N. Zaslavsky. "**Deep Learning and the Information Bottleneck Principle**." In 2015 IEEE Information Theory Workshop (ITW), 1–5, 2015.
4. Saxe, Andrew Michael, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. "**On the Information Bottleneck Theory of Deep Learning**." In International Conference on Learning Representations, 2018.