Information Bottleneck

Yangshuo He

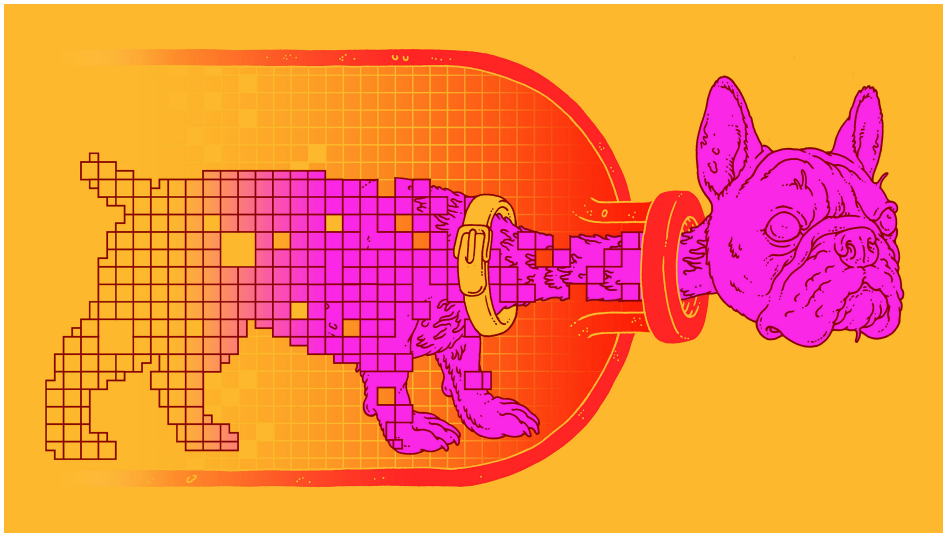
The University of Melbourne

Department of Electrical and Electronic Engineering

15/10/2025

Outline

- Background of the Information Bottleneck (IB) framework
- The IB cracks open the black box of deep learning
  - Two phases of the information dynamics during training
  - An information compression bounds
- Ongoing debates on the IB framework



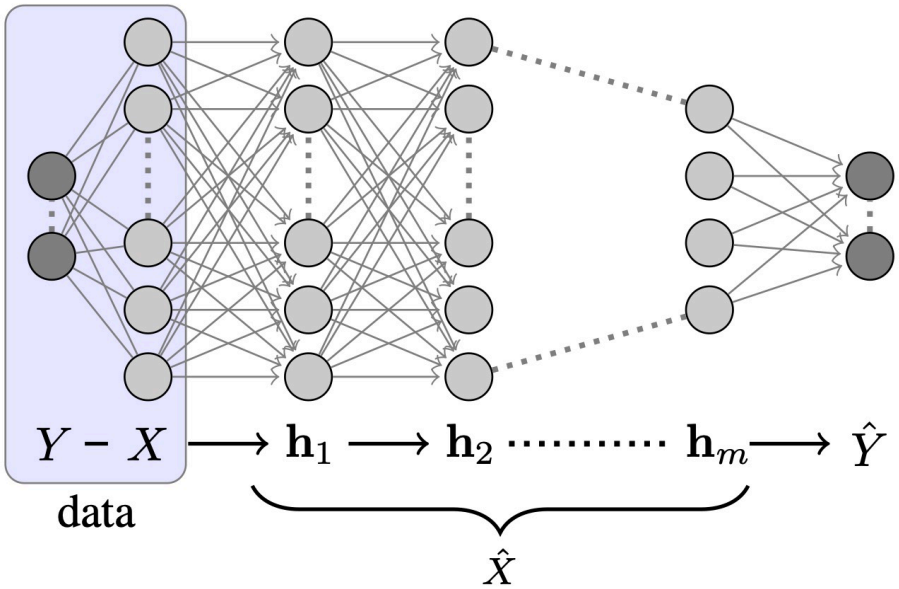
How does DNN work?

- Deep learning (DL) has shown its great potential in learning critical information for certain tasks
- The theoretical understanding of DL remains unsatisfied
- Information theory plays an important role in explaining the DL mechanism

IB combines 3 different ingredients

- Learning Theory: data-dependent but architecture-independent bound
- Information Theory: using the same notations and techniques
- Stochastic dynamics of the training: convergence of SGD

Review of DNN



- Y: target label
- X: input data
- $\mathbf{h}_i$: features
- $\hat{Y}$: prediction

- The input data goes through a series of transformations

$$\mathbf{h}_{i+1} = \sigma(W_i \mathbf{h}_i + b_i), \quad \mathbf{h}_0 = X \tag{1}$$

- The cascade network forms a **Markov chain**

$$Y \to X \to \hat{X} \to \hat{Y} \tag{2}$$

Information Theory Basics

- The KL divergence: for any two distributions p and q over $\mathcal{X}$:

$$D\left(p\|q\right) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \geq 0$$

- The mutual information: for any two r.v.s X and Y:

$$I(X;Y) = D\left(p(x,y)\|p(x)p(y)\right) = H(X) - H(X|Y)$$

- Data Processing Inequality (DPI): for any Markov chain $X \to Y \to Z$

$$I(X;Y) \geq I(X;Z)$$

The IB Method (Tishby, Pereira, and Bialek 1999)

- Standard compression

$$X \to \hat{X}$$

- A constraint through a distance function:

$$\mathcal{L}\left[p\left(\hat{x}|x\right)\right] = I(X;\hat{X}) + \beta \mathbb{E}\left[d\left(x,\hat{x}\right)\right]$$

- Relevance compression

$$Y \to X \to \hat{X}$$

- A constraint on the meaningful information:

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

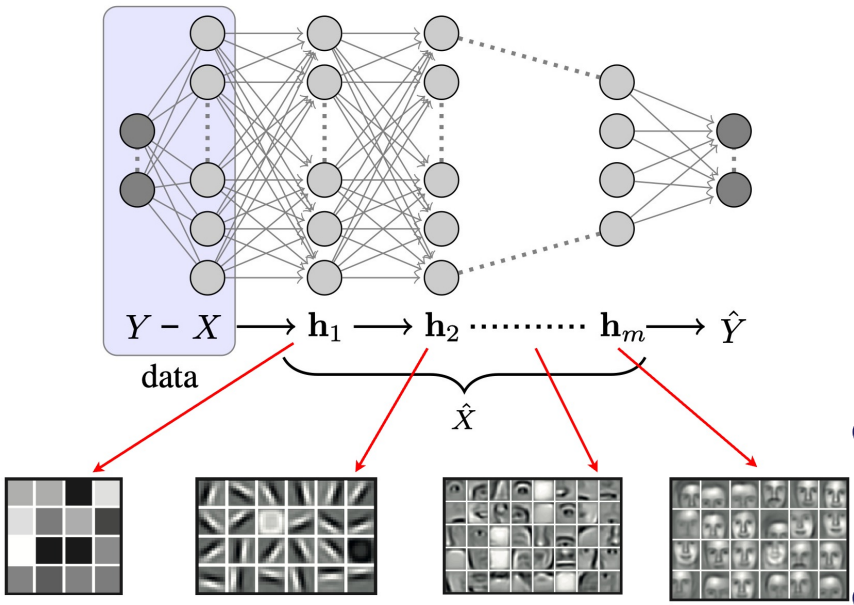
- Using the Lagrange multipliers method, the optimal solution is

$$p^*(\hat{x}|x) = \frac{p(\hat{x})}{Z(x,\beta)} \exp\left[-\beta D\left(p(y|x) \| p(y|\hat{x})\right)\right], \tag{3}$$

where $Z(x,\beta)$ is the normalization function

- Iteratively solved by Blauht-Arimoto algorithm

What do the DNN layers represent?



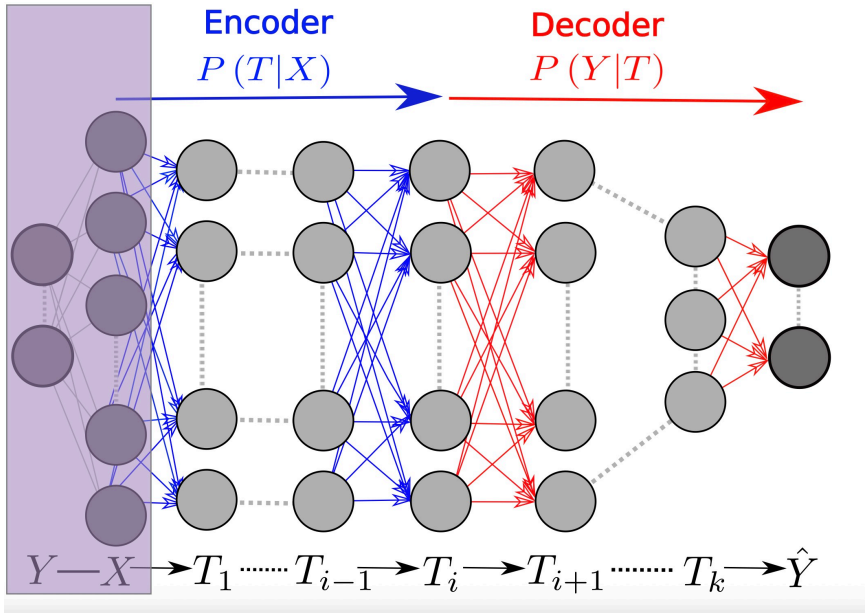
- Apply DPI on the cascade representations:

$$H(X) \geq I(X;\mathbf{h}_i) \geq I(X;\mathbf{h}_{i+1}) \geq \cdots \tag{4}$$

$$I(X;Y) \geq I(\mathbf{h}_i;Y) \geq I(\mathbf{h}_{i+1};Y) \geq \cdots \tag{5}$$

- Each layer gives a more refine representation
- These representations induce distinct partitions of input

Consider a DNN as a pair of encoder and decoder



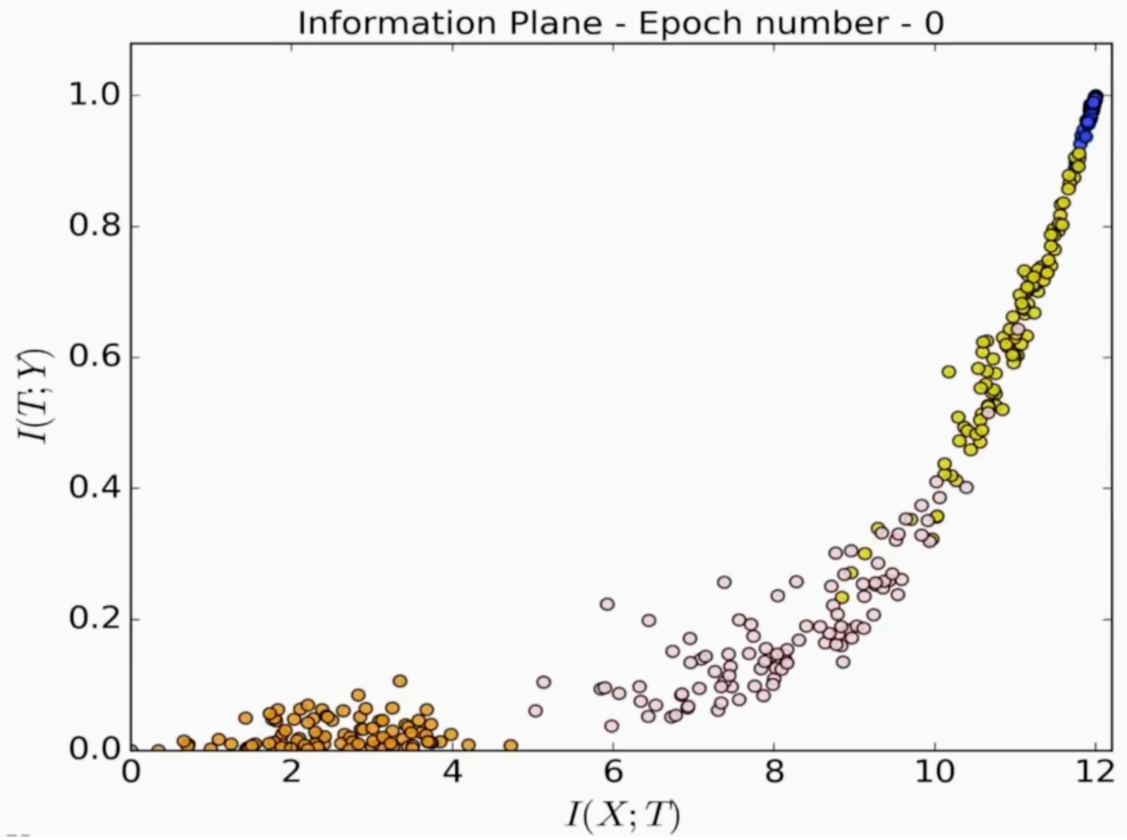
- A resemblance between DNN and relevance compression
- Learn a minimal sufficient statistics T (**bottleneck**): minimal description length $I(X;T)$ and maximal relevant information $I(T;Y)$

Statement (Tishby and Zaslavsky 2015)

The sample complexity of DNN is determined by the encoder MI $I(X;T)$, and the accuracy is determined by the decoder MI $I(T;Y)$

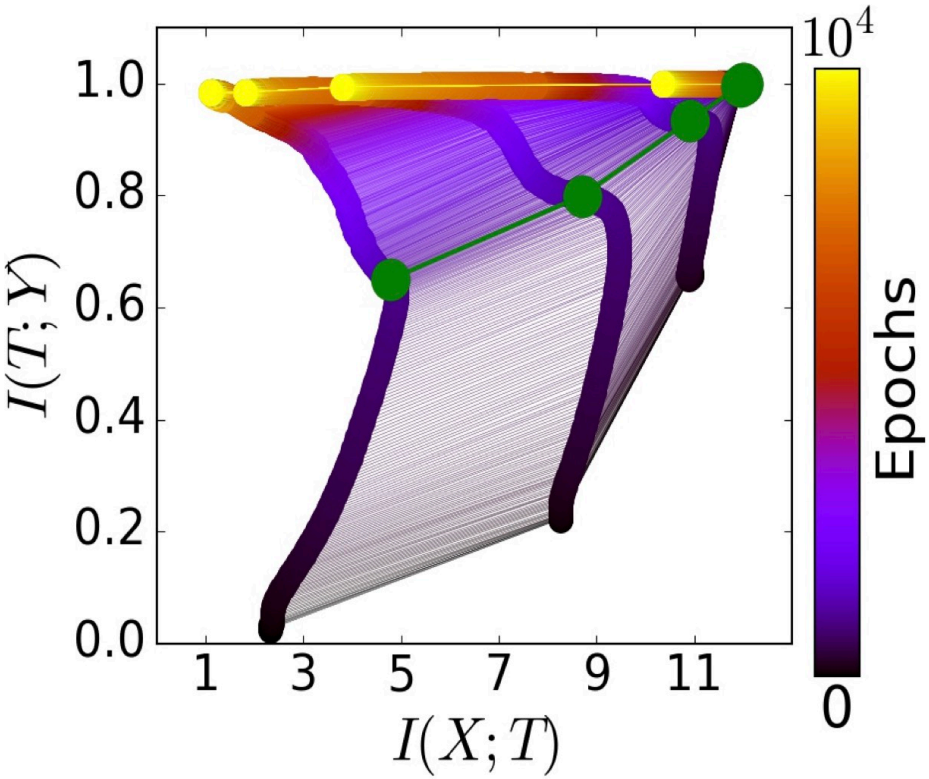
The Information Planes

- IB framework formulates a trade-off between $I(X;T)$ and $I(T;Y)$
- The set of all possible pairs of MIs for any encoder $p(T|X)$ is called the **information plane**



- Simulations demonstrate a two phases dynamics during training

Information Dynamics



- **Fitting Phase:** the layers increase the information on the labels
- **Compression Phase:** the layers lose irrelevant information about data

Rethinking Learning Theory (Intuitively)

- "Old" Generalization bound:

$$\epsilon^2 \leq \frac{\log |H| + \log \frac{2}{\delta}}{2m}$$

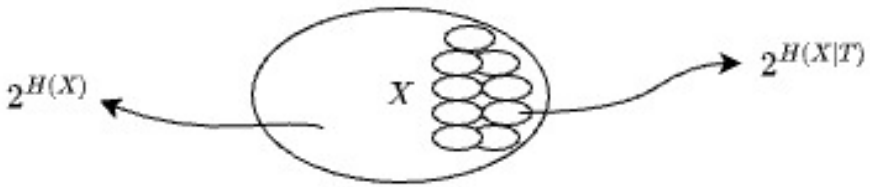
- ϵ: generalization error
- $|H|$: size of hypothesis space
- m: number of training samples
- δ: confidence
- **Don't work for DL!**

Rethinking Learning Theory (Intuitively)

- "Old" Generalization bound:

$$\epsilon^2 \leq \frac{\log |H| + \log \frac{2}{\delta}}{2m}$$

- ϵ: generalization error
- $|H|$: size of hypothesis space
- m: number of training samples
- δ: confidence
- **Don't work for DL!**



- Input Compression bound:
- $|H| \sim 2^{|X|} \to 2^{|T|}$
- Typical set and typical partition

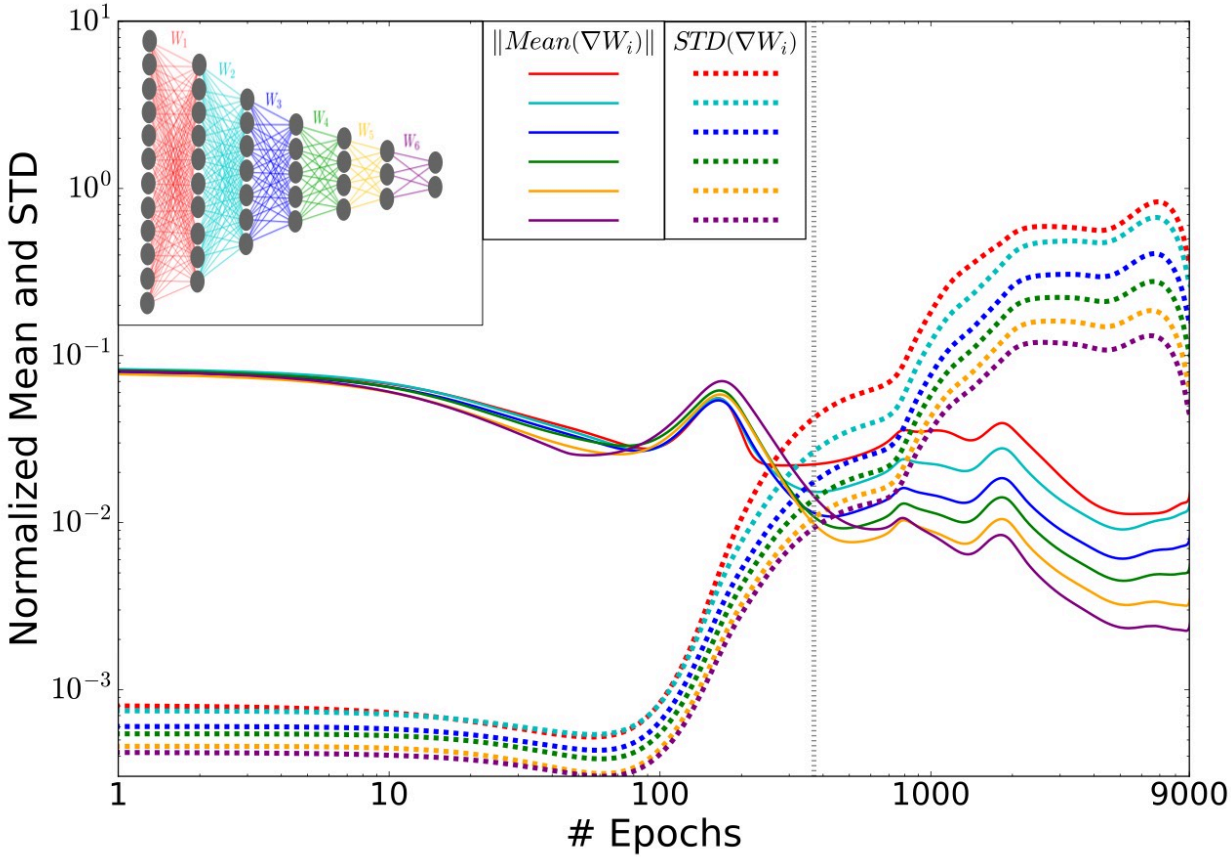
$$p(x_1, \cdots, x_n) \approx 2^{-nH(X)}$$
$$p(x_1, \cdots, x_n|T) \approx 2^{-nH(X|T)}$$

- The size of partition $|T| \sim 2^{I(X;T)}$

$$\epsilon^2 \leq \frac{2^{I(X;T)} + \log \frac{2}{\delta}}{2m}$$

- every bit of compression is like doubling the training samples.

Interesting Results

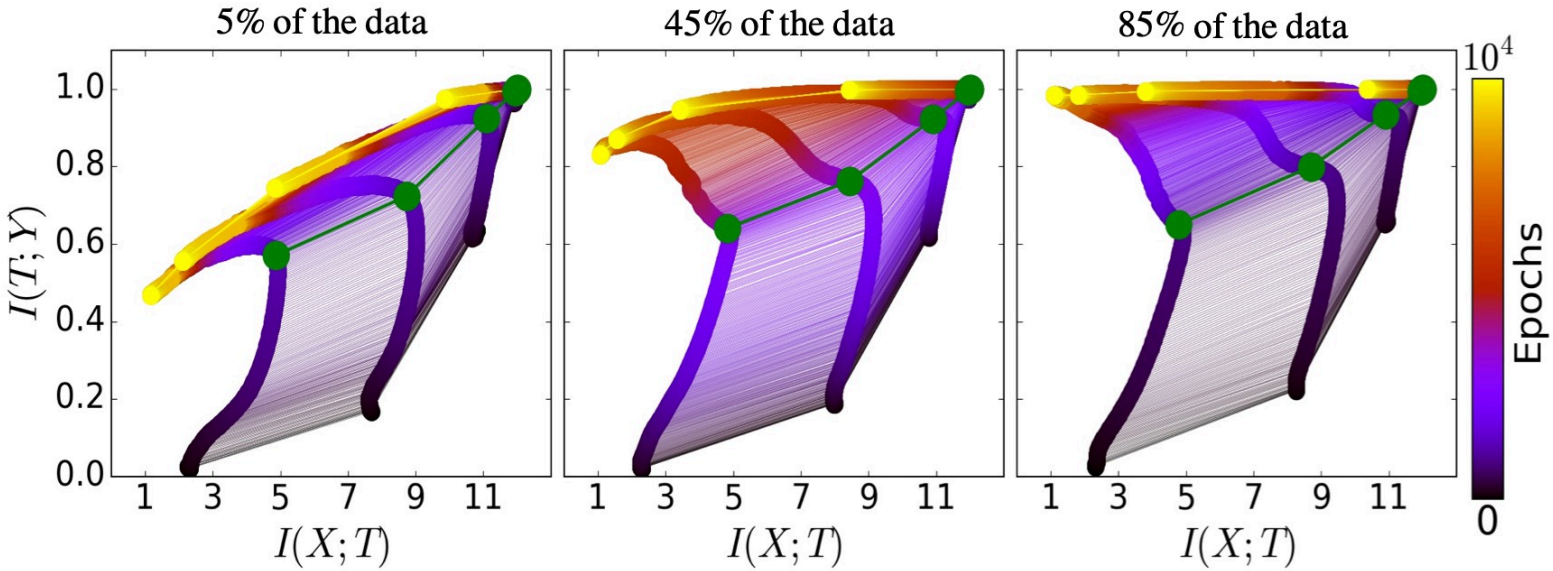


Two clear phases during training:

- High SNR phase: memorization (drift)
- Low SNR phase: forgetting (diffusion)

Claim: These two stochastic gradient phases correspond the fitting and compression phases

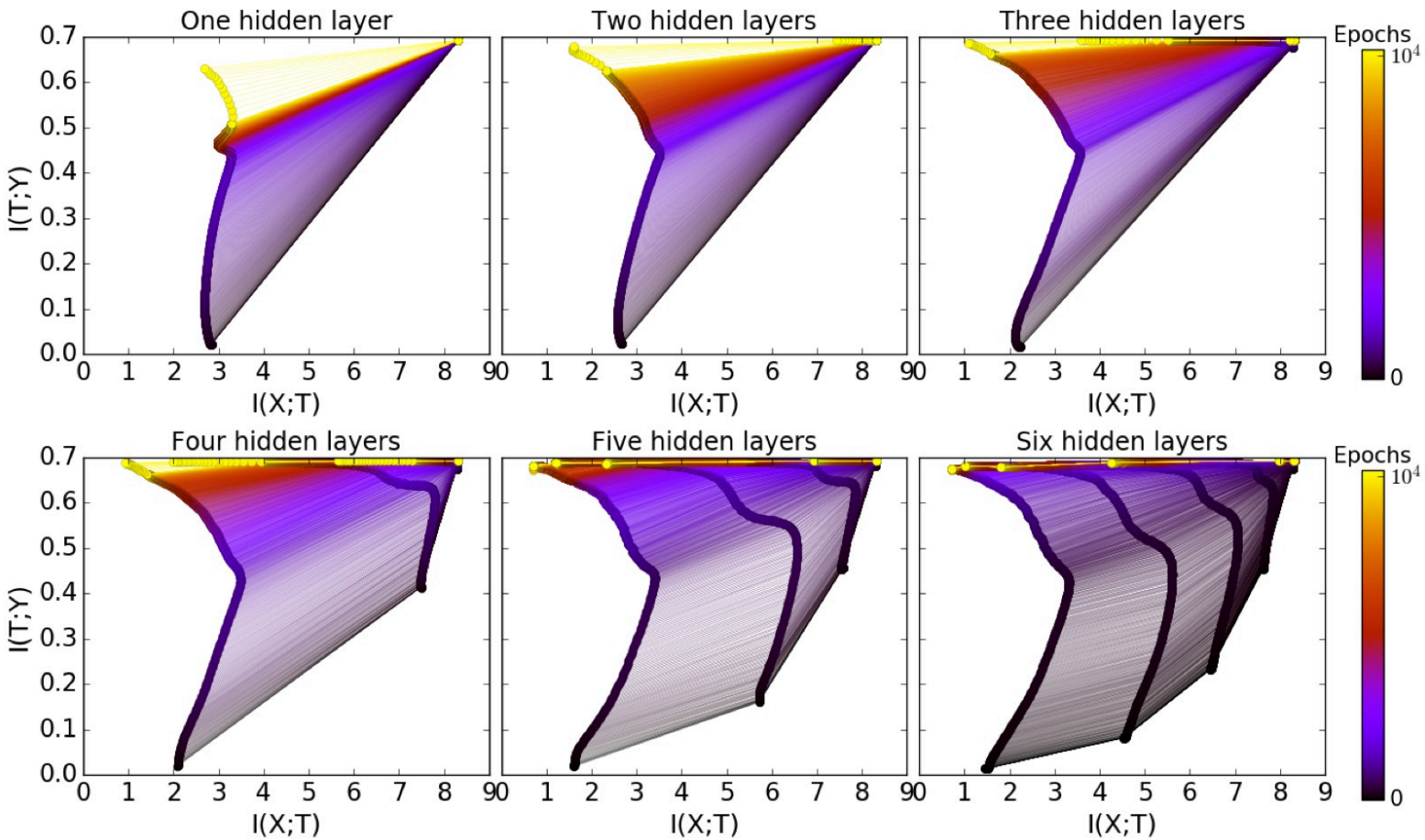
Interesting Results



The information dynamics of the layers, for different training samples

- The initial fitting phases are similar
- Lack of training data causes over-compression (over-fitting) in the compression phase

Interesting Results



The information dynamics of the layers, for different architectures

- Adding layers reduce the training epochs
- The compression is faster for deeper layers

Review of IB

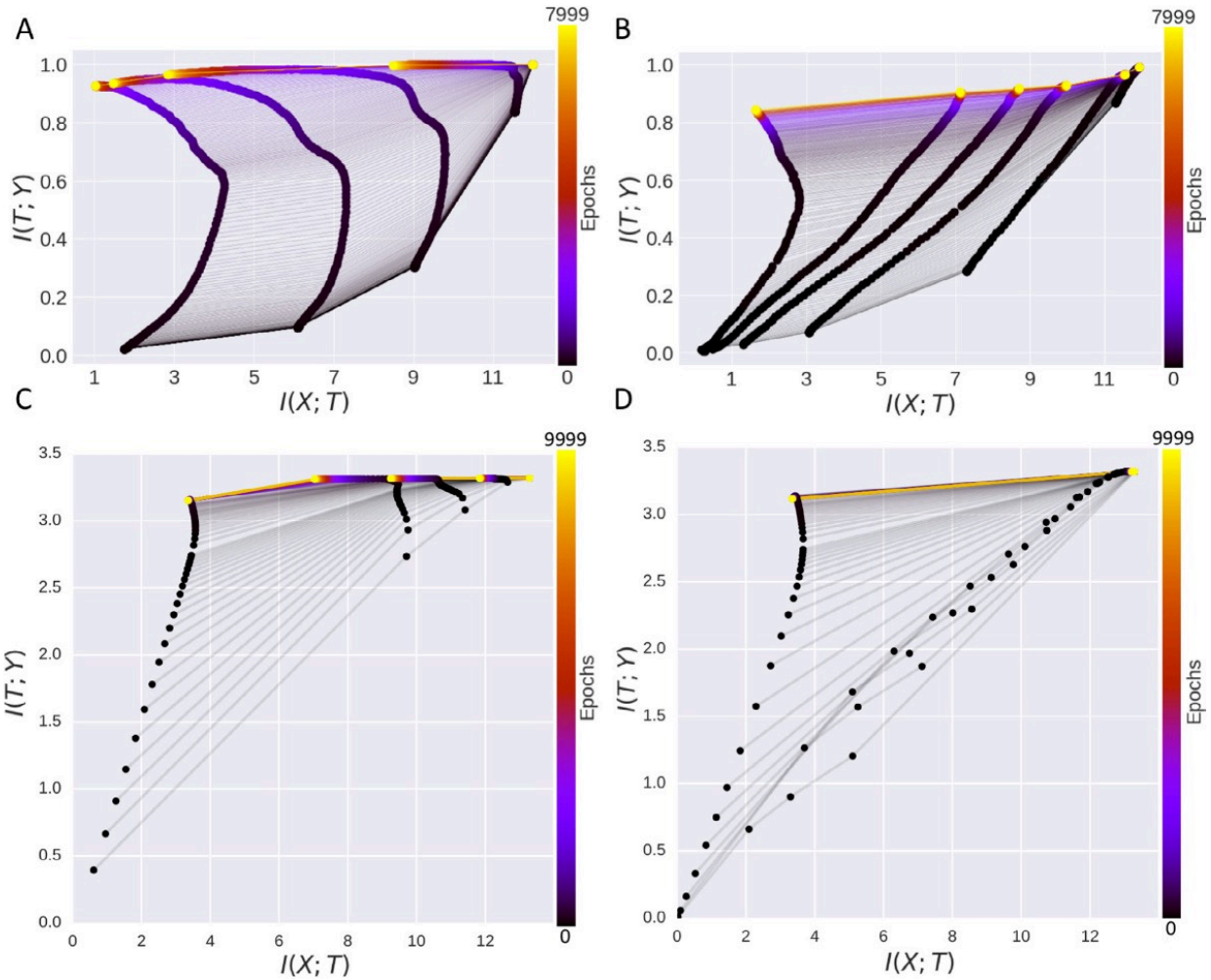
- DNN learns a bottleneck representation that optimizes the trade-off between compression and prediction
- Training DNN undergoes two distinct phases consisting of a fitting phase and a compresssion phase
- The compression phase contributes to the generalization performance of DNN
- The compression phase occurs due to the diffusion behaviour of SGD

Review of IB

- DNN learns a bottleneck representation that optimizes the trade-off between compression and prediction
- Training DNN undergoes two distinct phases consisting of a fitting phase and a compresssion phase
- The compression phase contributes to the generalization performance of DNN
- The compression phase occurs due to the diffusion behaviour of SGD

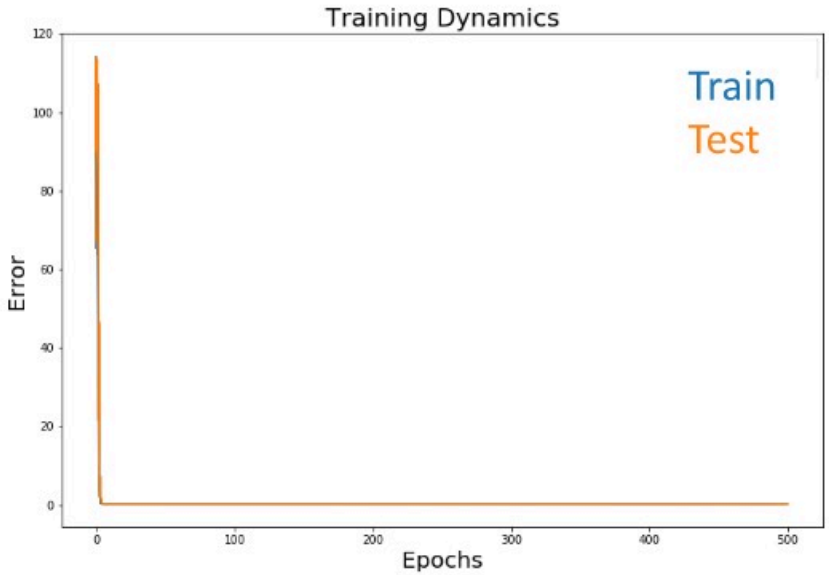
None of them hold true in the general case! (Saxe et al. 2019)

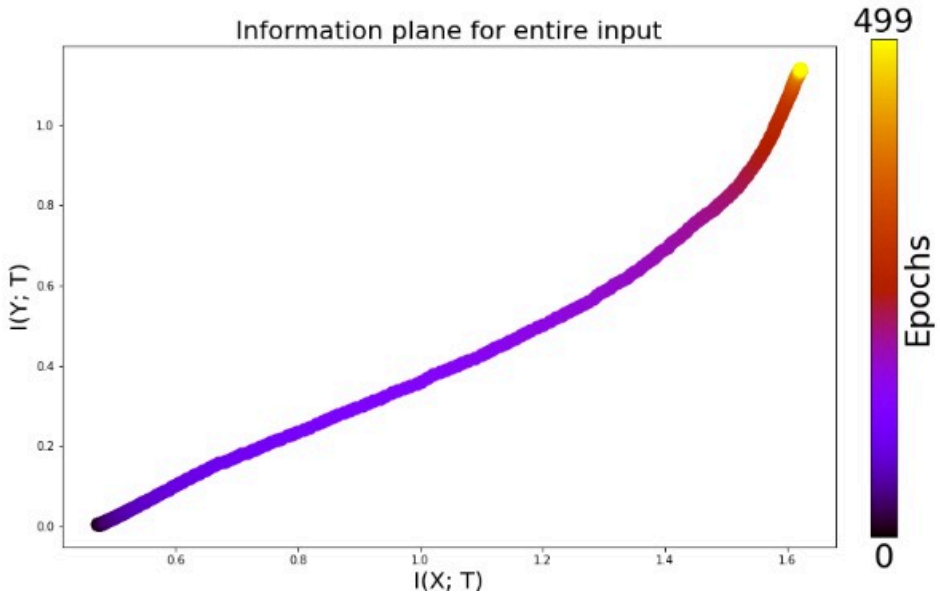
Challenge on the Compression Phase



- (A)(B): binary task
- (C)(D): MNIST dataset
- (A)(C): DNN with tanh
- (B)(D): DNN with ReLU
- No compression is observed except the final classification layer using sigmoid neurons
- **Compression arises from the saturation in non-linear activation**

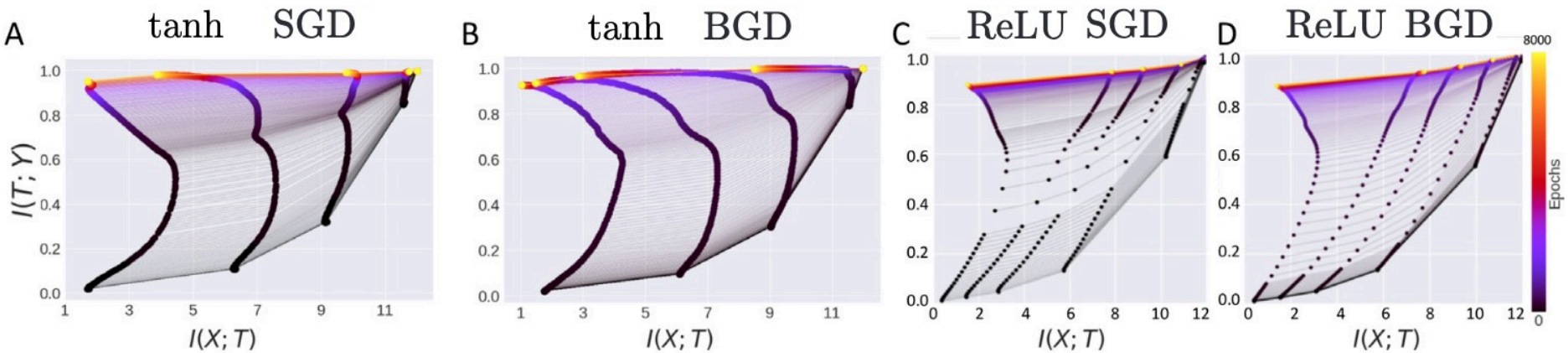
Challenge on Generalization





- Generalization and information plane in a deep linear network
- No compression is observed
- The network generalizes well

Challenge on Diffusion Behaviour



- SGD: stochastic gradient descent; BGD: full batch gradient descent (no randomness)
- The randomness in the training does not contribute to compression of information

Reference I


Tishby, Naftali, Fernando C. Pereira, and William Bialek (1999). "The information bottleneck method". In: *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377. URL: https://arxiv.org/abs/physics/0004057.


Tishby, Naftali and Noga Zaslavsky (2015). "Deep learning and the information bottleneck principle". In: *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5. DOI: 10.1109/ITW.2015.7133169.


Saxe, Andrew M et al. (Dec. 2019). "On the information bottleneck theory of deep learning". In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12, p. 124020. DOI: 10.1088/1742-5468/ab3985. URL: https://doi.org/10.1088/1742-5468/ab3985.