ON THE INFORMATION BOTTLENECK THEORY OF DEEP LEARNING

Andrew M. Saxe, Yamini Bansal, Joel Dapello, Madhu Advani
Harvard University
{asaxe,madvani}@fas.harvard.edu,{ybansal,dapello}@g.harvard.edu

Artemy Kolchinsky, Brendan D. Tracey
Santa Fe Institute
{artemyk,tracey.brendan}@gmail.com

David D. Cox
Harvard University
MIT-IBM Watson AI Lab
davidcox@fas.harvard.edu
david.d.cox@ibm.com

Accepted at ICLR 2019
Scores: 7 7 6

Information Bottleneck

1. Deep networks undergo two distinct phases consisting of an initial fitting phase and a subsequent compression phase.

2. The compression phase is causally related to the excellent generalization performance of deep networks.

3. The compression phase occurs due to the diffusion-like behavior of stochastic gradient descent.

R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Information Bottleneck

1. Deep networks undergo two distinct phases consisting of an initial fitting phase and a subsequent compression phase.

2. The compression phase is causally related to the excellent generalization performance of deep networks.

3. The compression phase occurs due to the diffusion-like behavior of stochastic gradient descent.

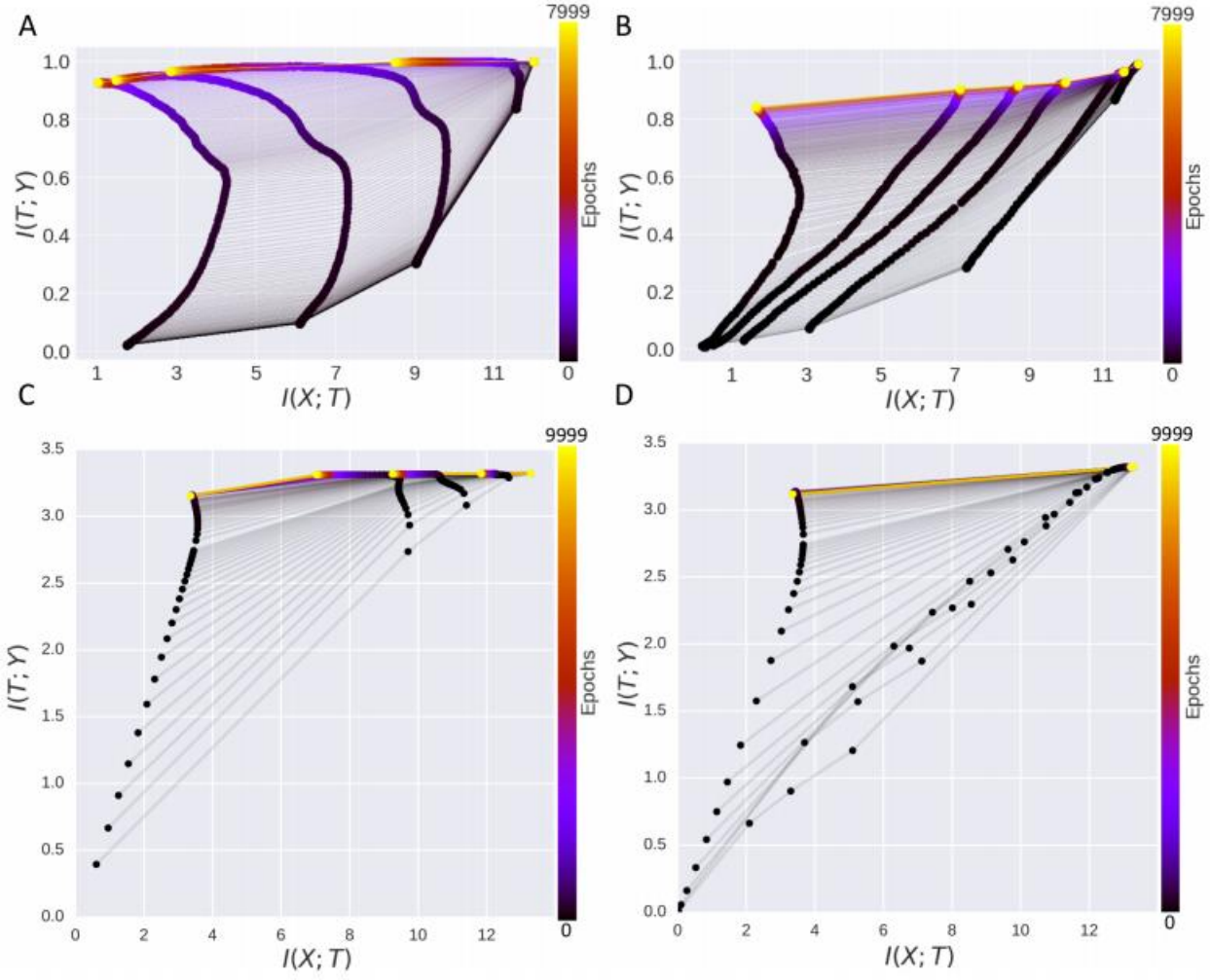
This paper: none of these claims hold true in the general case.

R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Point 1: Two Phases

- The information plane trajectory is a function of the neural nonlinearity employed:
  - Double-sided saturating nonlinearities like tanh yield a compression phase as neural activations enter the saturation regime.
  - Linear activation functions and single-sided saturating nonlinearities like the widely used ReLU in fact do not.

Point 1: Two Phases



A: tanh neural network layers show compression.

B: ReLU neural network layers show no compression.

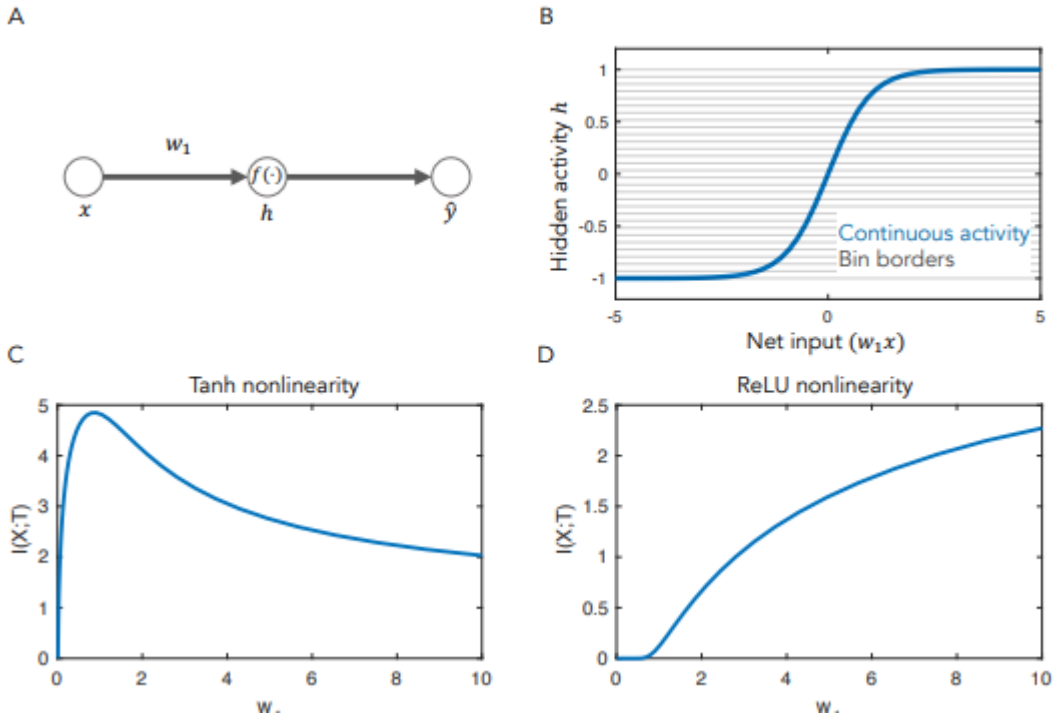
C: large network on the MNIST dataset, tanh networks compress.

D: large network on the MNIST dataset, ReLU networks do not compress.

Point 1: Two Phases

Explicitly calculate the mutual information for a simple model:

$$
\begin{aligned}
I(T;X) &= H(T) - H(T|X) \\
&= H(T) \\
&= -\sum_{i=1}^{N} p_i \log p_i
\end{aligned}
\qquad
p_i = P(X \geq f^{-1}(b_i)/w_1 \text{ and } X < f^{-1}(b_{i+1})/w_1),
$$



A: a simple three neuron nonlinear network.

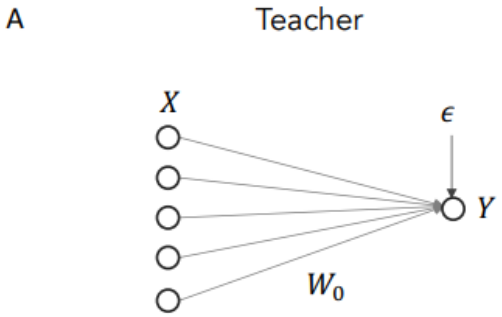
B: the activity h is binned into a discrete variable.

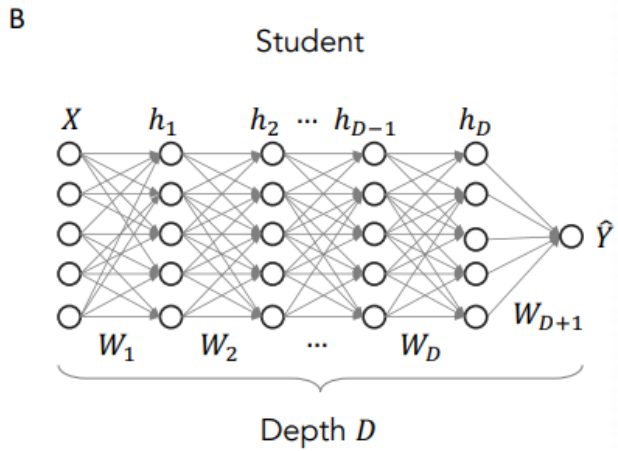
C: MI as a function of w_1 for tanh nonlinearity.

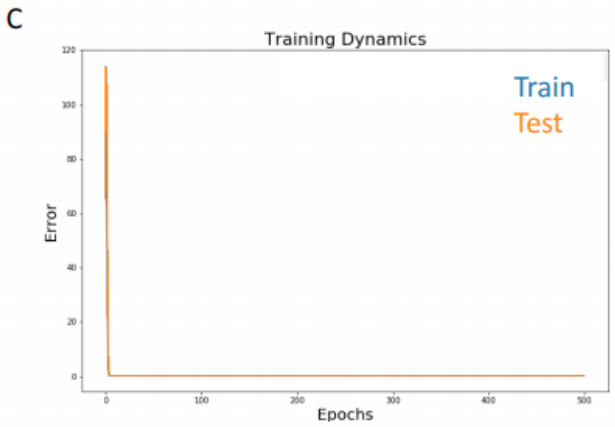
D: MI as a function of w_1 for ReLU nonlinearity.

Point 2: Compression and Generalization

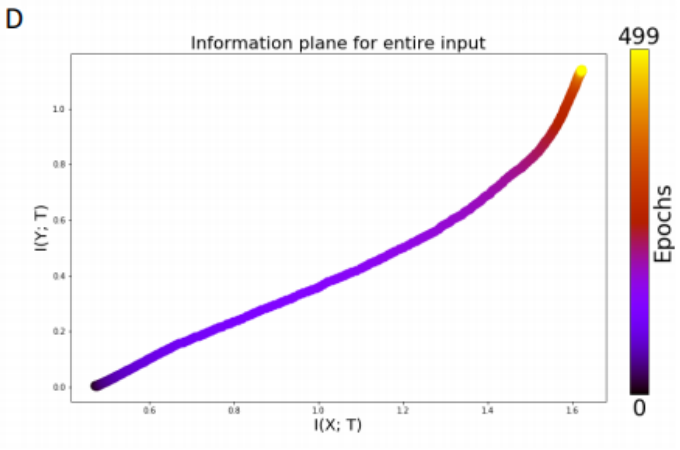
- Train simple linear networks in a student-teacher setup:
  - A linear teacher network generates a dataset.
  - A deep linear student network is trained on the dataset for generalization.
    - ✓ Tractable: $I(T;X) = \log|\bar{W}\bar{W}^T + \sigma_{MI}^2 I_{N_h}| - \log|\sigma_{MI}^2 I_{N_h}|$





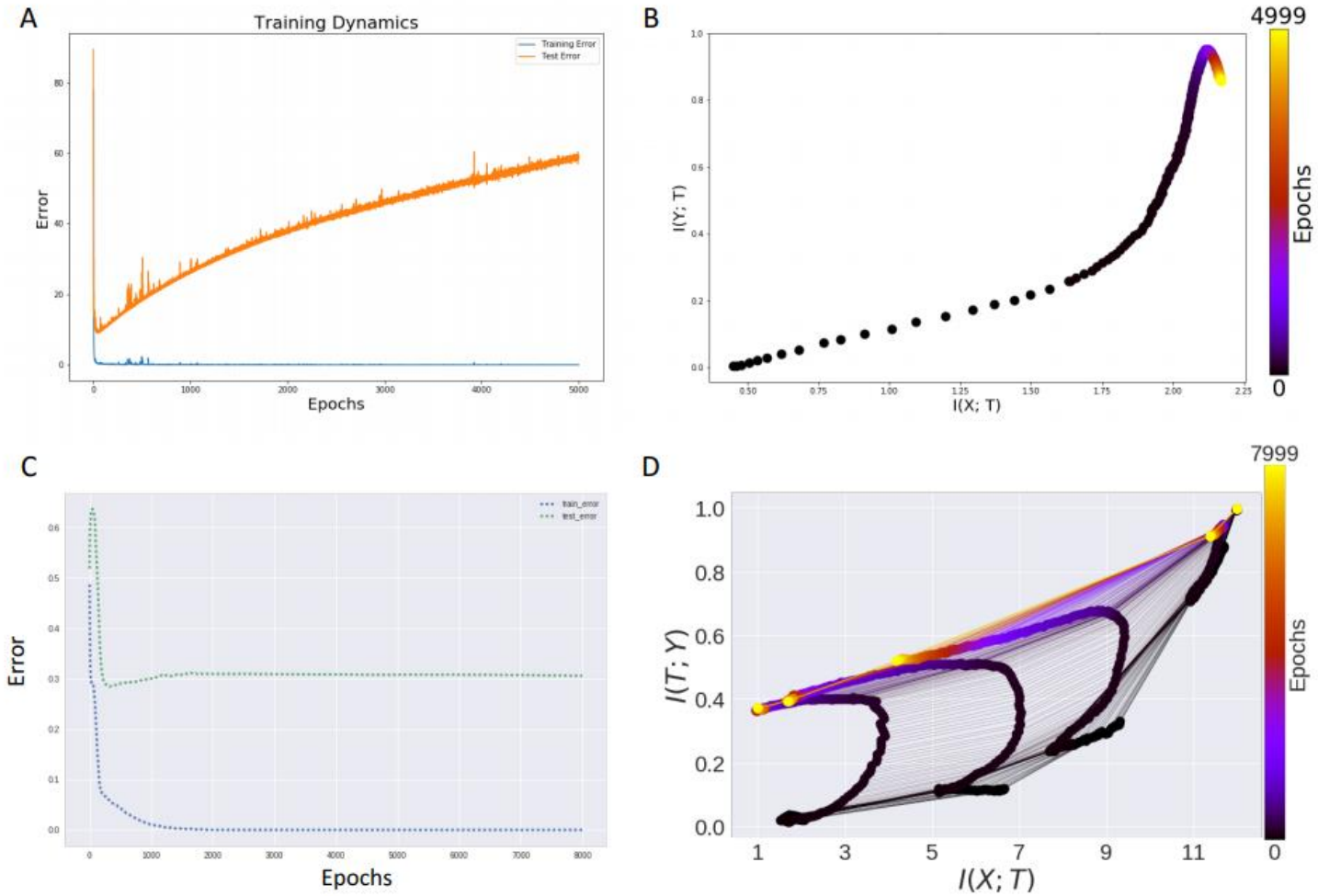


C: Training and testing error, show no overfitting.



D: Information plan dynamics show no compression.

Point 2: Compression and Generalization



A, B: deep linear network with SGD
not compress, not generalize well (overfitting).

C, D: nonlinear tanh network

compress, but not generalize well (overfitting).

Point 3: Compression and Stochasticity

- SGD comprises of “drift” and “diffusion” phase, compression occurs at the second phase.
- Batch Gradient Descent (BGD) uses the full training dataset and has no randomness or diffusion-like behavior in its updates.

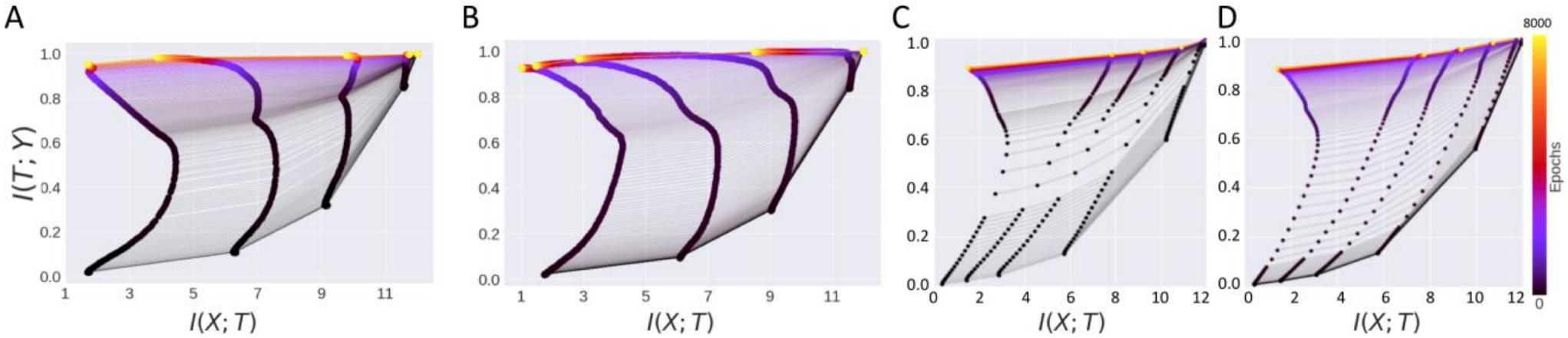


Figure 5: Stochastic training and the information plane. (A) tanh network trained with SGD. (B) tanh network trained with BGD. (C) ReLU network trained with SGD. (D) ReLU network trained with BGD. Both random and non-random training procedures show similar information plane dynamics.

Response From Tishby&Shwartz-Ziv

This “paper” attacks our work through the following flawed and misleading statements:

1. That the compression of the representation (reducing I(T:X)) is due to the saturated non-linearity and is not appear with other non-linearity (RelU’s in particular).

The authors don’t know how to estimate mutual information correctly. When properly done, there essentially the same fitting and compression phases with RelU’s and any other network we examined:

2. “that there is no evident causal connection between compression and generalization”

We rigorously proved that compression leads to dramatic improvement in generalization, providing that the partitions remained homogenous to the label probability. In fact we argue that any bit of representation compression (under these conditions) is effective as doubling the size of the training data! Here is the sketch of our proof as given in our presentations:

3. “that the compression is unrelated to the noisy (low SNR) phase of the gradients”, as we claim.

Moreover, when changing the min-batch size (from 32 to 4000) both transitions move together in perfect linear relationship (left). In fact we show (right) that the full batch case (BGD of the “paper”) lies on the same line (green point) which suggests that the reported compression here is exactly the same phenomena, for much weaker gradient noise (as we claimed).

We believe these facts nullify the arguments given in this “paper” all together.