UVA CS 6316: Machine Learning : 2019 Fall
Course Project: Deep2Reproduce @
https://github.com/qiyanjun/deep2reproduce/tree/master/2019Fall

ON THE INFORMATION BOTTLENECK THEORY OF DEEP LEARNING

Reproduced by: Siyuan Liu Yusheng Jiang Zhidan Luo

Dec. 2019

Background

From information bottleneck theory of deep learning, we get three specific claims:

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.

Motivation

- Our paper shows that none of these claims hold true in the general case.

- Deep neural networks are the tool of choice for real-world tasks ranging from visual object recognition to unsupervised learning and reinforcement learning. These practical successes have spawned many attempts to explain the performance of deep learning systems.

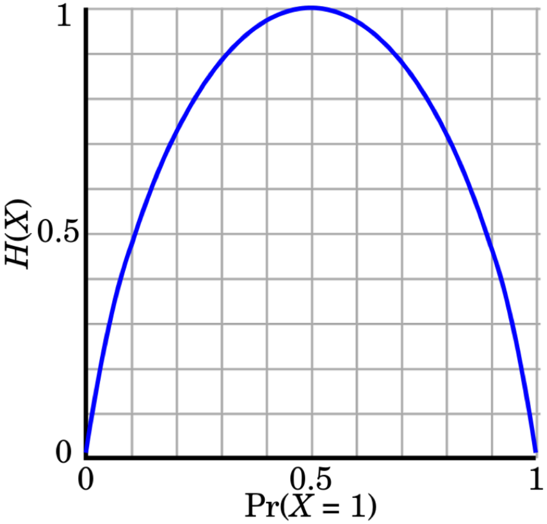
Related Work

Information theory

Based on the probability mass function of each source symbol to be communicated, the Shannon entropy H, in units of bits (per symbol), is given by

$$H = -\sum_i p_i \log_2 (p_i)$$

where pi is the probability of occurrence of the i-th possible value of the source symbol.



Related Work

Deep learning

There is a close connection between machine learning and compression: a system that predicts the posterior probabilities of a sequence given its entire history can be used for optimal data compression (by using arithmetic coding on the output distribution) while an optimal compressor can be used for prediction (by finding the symbol that compresses best, given the previous history). This equivalence has been used as a justification for using data compression as a benchmark for "general intelligence."

Claim / Target Task

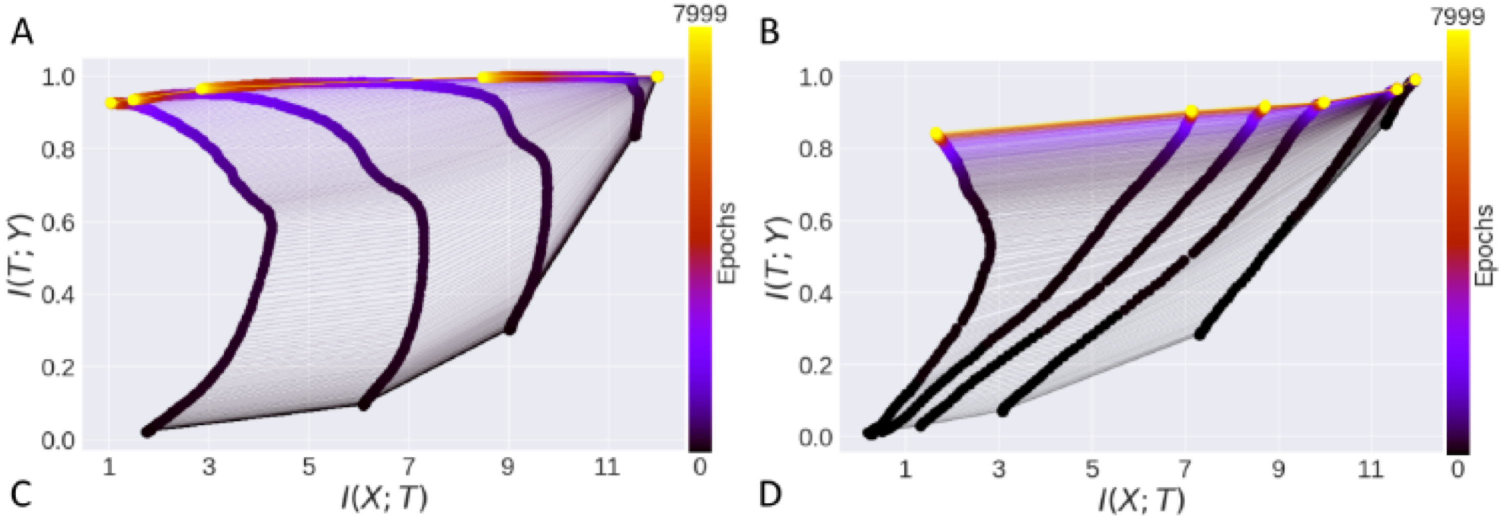
1. The information plane trajectory is predominantly influenced by the choice of neural activation functions.

2. There is no evident causal connection between compression and generalization.

3. The compression phase, when it exists, does not arise from stochasticity in training.

4. When an input domain consists of a subset of task-relevant and task-irrelevant information, hidden representations do compress the task-irrelevant information. Further, the compression happens concurrently with the fitting process rather than during a subsequent compression period.

An Intuitive Figure Showing WHY Claim



Proposed Solution

1. Use ReLU instead of the activation function tanh to observe the information plane dynamics.

2. Explore the connection between generalization and compression.

3. Research whether stochasticity is important for compression in training.

4. Research on the extent to which information related to tasks is compressed.

Implementation

- Replicate of the result reported by Shwartz-Ziv & Tishby for networks with the tanh nonlinearity.

- Modify the code to train deep networks using rectified linear activation functions ($f(x) = \max(0, x)$).

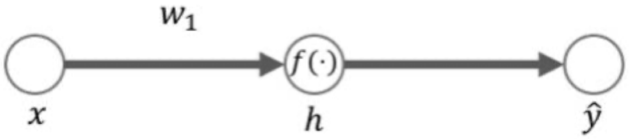
- Train networks on the MNIST dataset and computed mutual information.

- Develop a minimal model to understand the impact of neural nonlinearity on the mutual information dynamics.

Data Summary

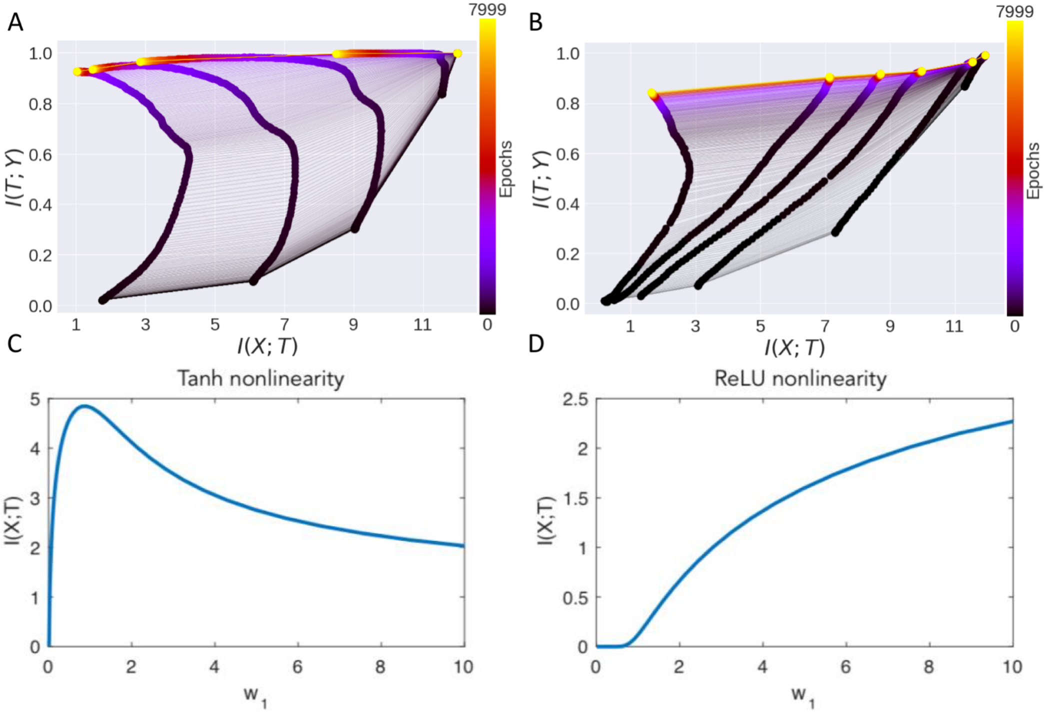
- **Model in replication:** a neural network with 7 fully connected hidden layers of width 12-10-7-5-4-3-2 is trained with SGD to produce a binary classification from a 12-dimensional input. 256 randomly selected samples per batch are used.

- **Minimal model:** a scalar Gaussian input distribution $X \sim \mathcal{N}(0,1)$, which is fed through the scalar first layer weight w_1, and passed through a neural nonlinearity $f(\cdot)$, yielding the hidden unit activity $h = f(w_1 X)$.



Experimental Results

I: Compression and Neural Nonlinearities



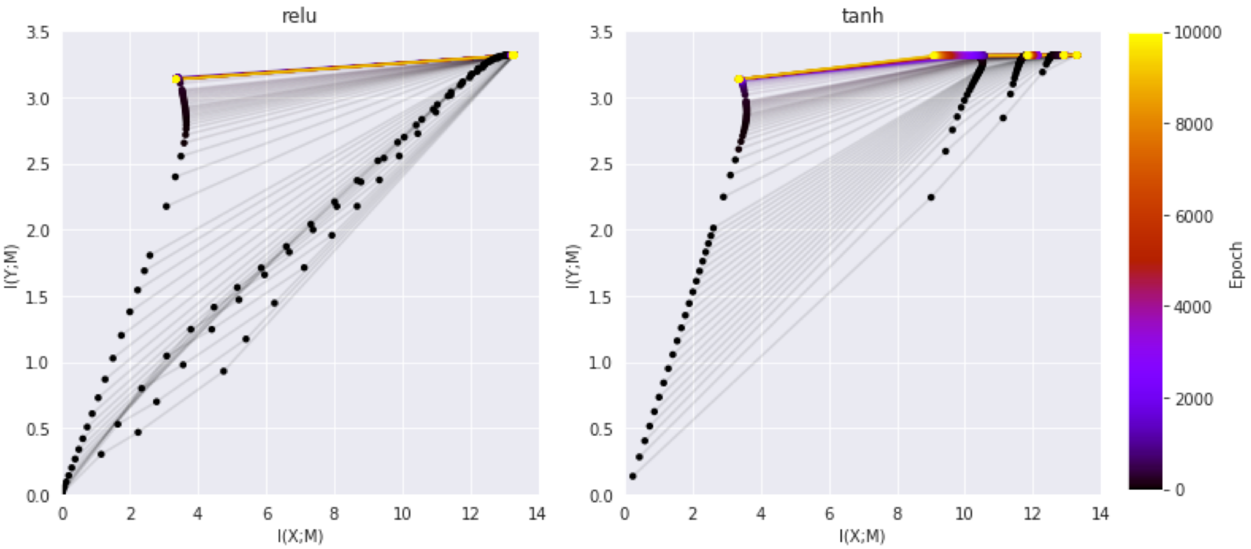
Except for the final layer of 2 sigmoidal neurons:

- The mutual information with the input monotonically increases in all ReLU layers
- No compression phase is visible in the ReLU layers any more.

Experimental Results

I: Compression and Neural Nonlinearities

Our results:



MNIST Dataset:

$$I(T;X) = -\frac{1}{P}\sum_i \log \frac{1}{P}\sum_j \exp\left(-\frac{1}{2}\frac{\|h_i - h_j\|_2^2}{\sigma^2}\right)$$

- 60000 samples each with 784 features
- 10 categories output
- 7 fully connected hidden layers of width 1024-25-25-20-10
- SGD with batch size of 128

Experimental Analysis

I: Compression and Neural Nonlinearities

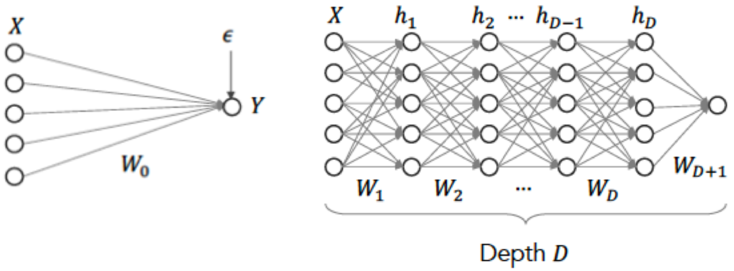
- The choice of nonlinearity substantively affects the dynamics in the information plane.

- The double-sided saturation of tanh is the key to the original result.

Implementation

II: Information Plane Dynamics in Deep Linear Networks

- **Linear networks:** exploit recent results on the generalization dynamics in simple linear. The activation function of each neuron is simply $f(u) = u$.



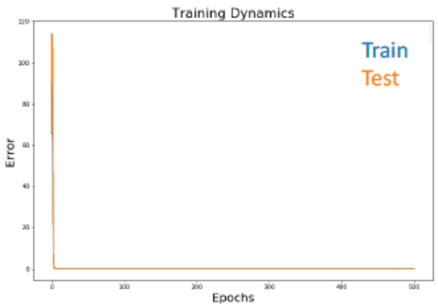
- Over-train linear networks to observe the information plane dynamics.

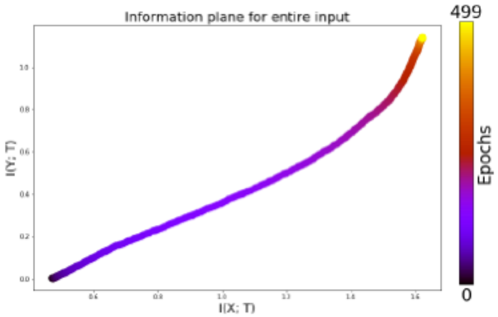
- **Nonlinear networks:** train tanh networks in the same previous setting with 30% of the data.

Experimental Results

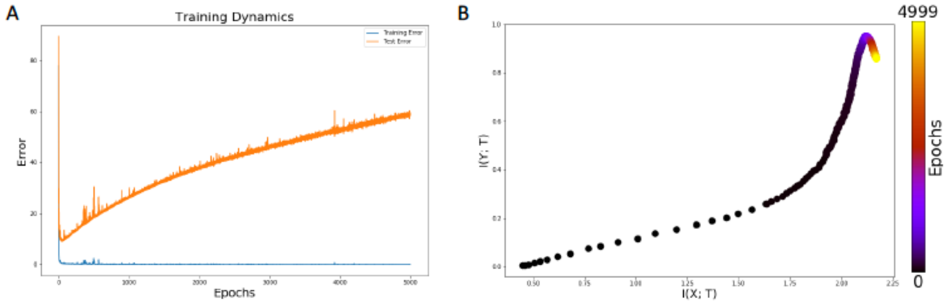
II: Information Plane Dynamics in Deep Linear Networks

Linear networks:





Over-training:

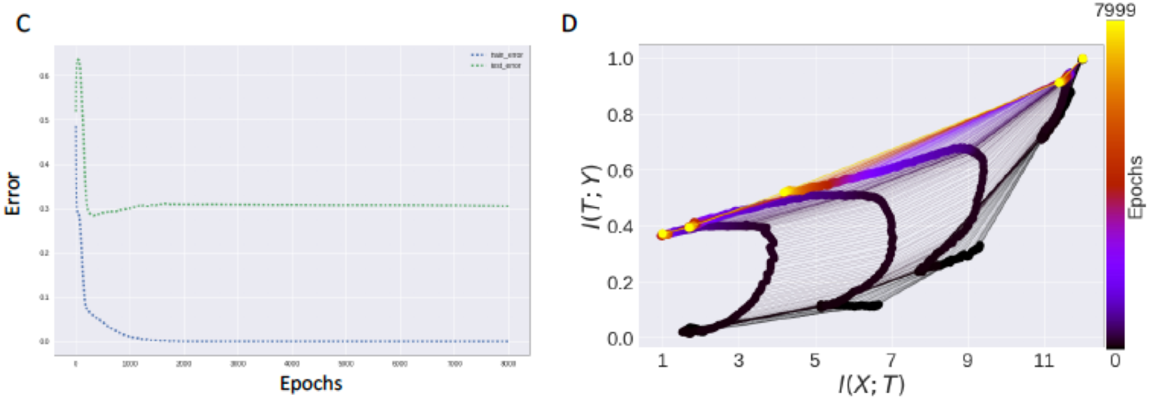


- No compression is seen in both information planes.
- Both networks exhibit similar information dynamics, but yield different generalization performance.

Experimental Results

II: Information Plane Dynamics in Deep Linear Networks

Nonlinear networks:



- The tanh networks show substantial compression, despite exhibiting modest overtraining.

- However, it did not show a good generalization performance.

Experimental Analysis

II: Information Plane Dynamics in Deep Linear Networks

- This establishes dissociation between behavior in the information plane and generalization dynamics: networks that compress may or may not generalize well, and networks that do not compress may or may not generalize well.
- Generalization performance can be acceptable without any compression phase.
- Networks with similar dynamics on the information plane may have different generalization performance.

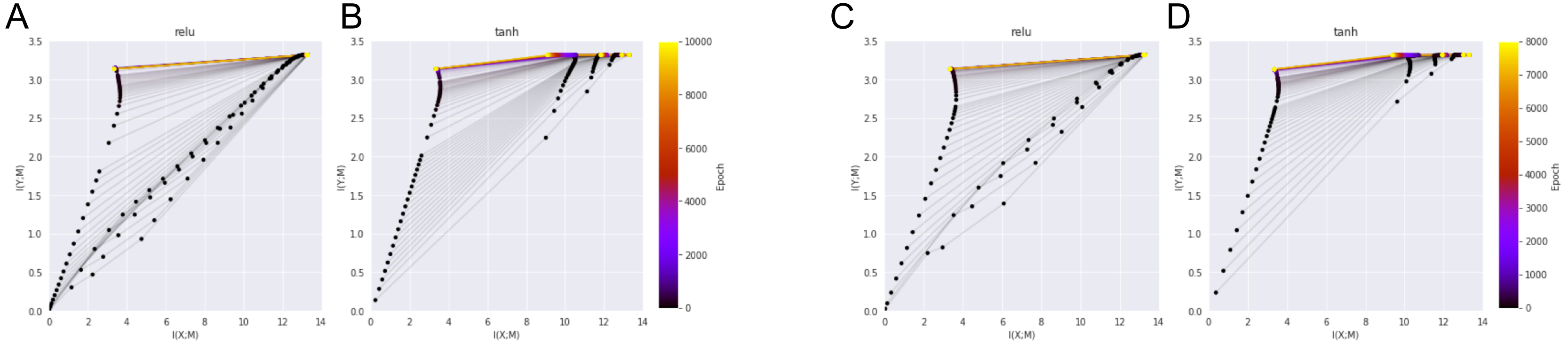
Implementation

- **Stochastic gradient descent:** it learns from a fixed-size dataset, and updates weights by repeatedly sampling a single example from the dataset and calculating the gradient of the error with respect to that single sample.

- **Batch gradient descent:** it learns from a fixed-size dataset, and updates weights using the gradient of the total error across all examples. **Crucially**, it has no randomness or diffusion-like behavior in its updates.

Experimental Results

III: Compression in BGD and SGD



(A) ReLU network trained with SGD

(B) tanh network trained with SGD

(C) ReLU network trained with BGD

(D) tanh network trained with BGD

- Both random and non-random training procedures show similar information plane dynamics.
- It shows robust compression in tanh networks for both methods.

Experimental Analysis

III: Compression in BGD and SGD

- Randomness in the training process does not appear to contribute substantially to compression of information about the input.

- This finding is consistent with the view presented in Section I that compression arises predominantly from the double saturating nonlinearity.

Implementation

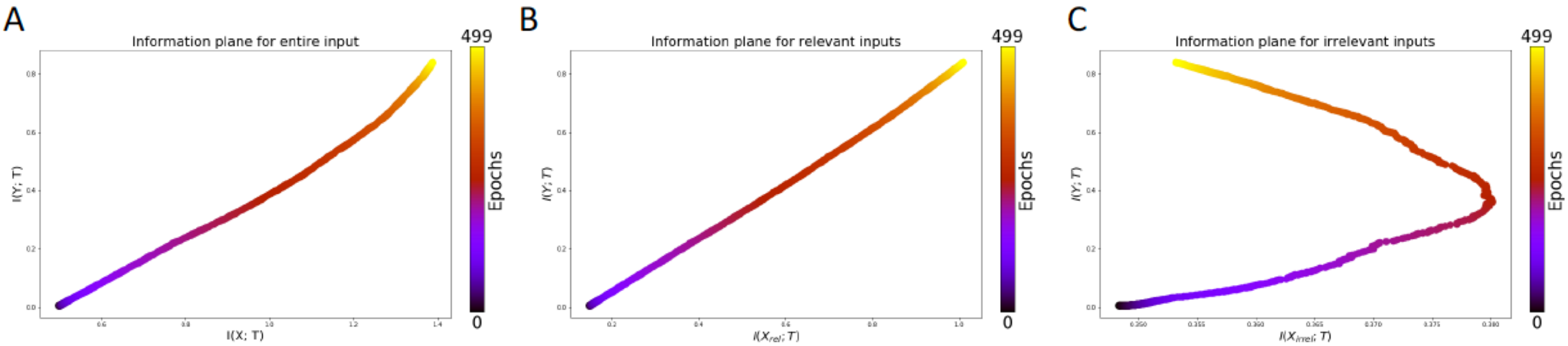
IV: Simultaneous Fitting and Compression

- The input X is divided into a set of task-relevant inputs X_{rel} and a set of task-irrelevant inputs X_{irrel}, and alter network so that the weights to the task-irrelevant inputs are all zero.

- The inputs X_{irrel} contribute only noise, while the X_{rel} contain signal. We then calculate the information plane dynamics for the whole layer, and for the task-relevant and task-irrelevant inputs separately.

Experimental Results and Analysis

IV: Simultaneous Fitting and Compression



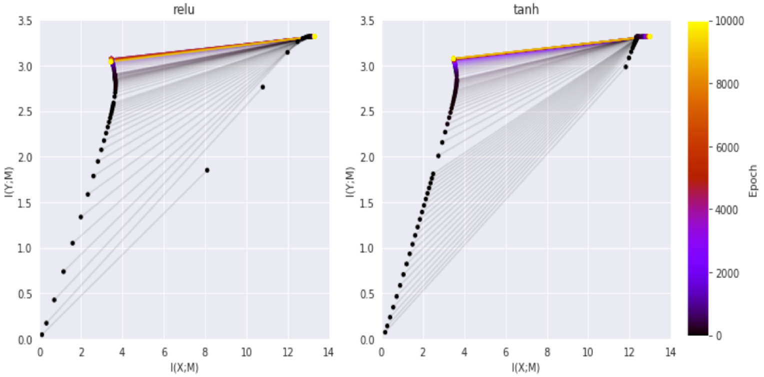
(A) A task with a large task-irrelevant subspace in the input
(B) The information with the task-relevant subspace
(C) The information with the task-irrelevant subspace

Simultaneous fitting and compression:

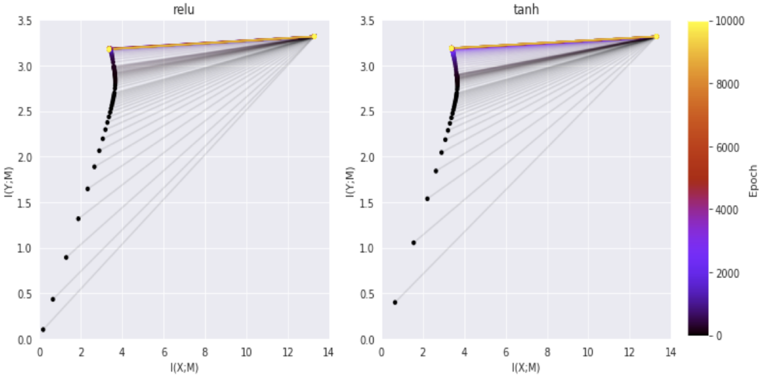
- the overall dynamics show no compression phase
- task-relevant subspace increases robustly over training
- task-irrelevant subspace does compress over the course of training

Our Comments

Architecture: 20-10

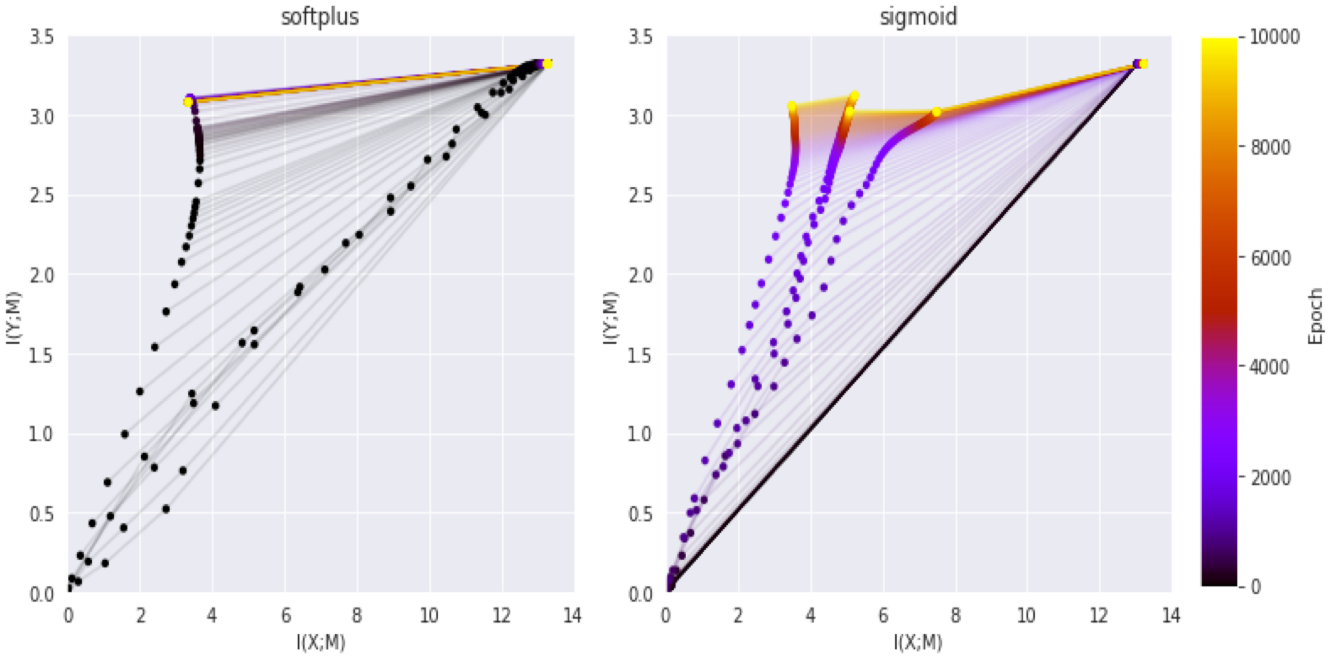


Architecture:512-10



- When the network architecture is simple, the dynamics of information plane show no compression phase for both ReLU and tanh activation function.

Our Comments



- Sigmoid is a double-sided activation function, but the dynamics of information plane doesn't show compression. We doubt that the compression is due to the double-saturating activation function.

Conclusion and Future Work

Conclusion:

- To prove dynamical compression in information plane maybe not a general characteristic of deep network learning
- Double-saturating nonlinearities contribute to compression, but single-sided saturating nonlinearities such as ReLUs are generally not helpful to compress
- Doubt with the causal relationship between compression and generalization

Future Work

- Sigmoid and tanh are both double-sided saturation function. But they have different results. What properties decide the compression phase?

Responsibilities

- Yusheng Jiang: Section 3 and 4
- Zhidan Luo: Section 1 and 2
- Siyuan Liu: Section 2 and 3

References

- R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810, 2017.
- N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In IEEE Information Theory Workshop, 2015.
- N. Tishby, F.C. Pereira, and W. Bialek. The information bottleneck method. Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing, pp. 368–377, 1999.
- https://openreview.net/forum?id=ry_WPG-A-

References

Thanks!