Deep Mutual Information

Andrew Stirn[1], Robert Kwiatkowski[1], Iddo Drori[1,2,3]

[1] Columbia University, Department of Computer Science
[2] NYU, Center for Data Science, Courant Institute of Mathematical Sciences
[3] NYU, Tandon School of Engineering, Department of Computer Sciences

Abstract

We compute the mutual information between layers of deep neural networks used in supervised, unsupervised, and reinforcement learning. Our work compares between pairs of neural networks trained with different optimizers, regularization methods, and architectures, while all other factors being equal. We use the mutual information plane for explaining the generalization of deep neural networks. We collect mutual information measurements over combinations of data sets, optimizers, network architectures, and regularization methods for visualizing which permutations work and which configurations collapse. Our work extends mutual information beyond simple architectures and datasets by computing the deep mutual information of generative adversarial network (GAN) discriminators and Alpha Zero network players.

Motivation

Recent work [1] employed the mutual information to examine the learning characteristics of deep neural networks (DNN). These networks have widely been treated as black boxes despite their prolific use and their often superhuman performance. In this work, the authors studied the learning behavior of DNN's when trained using stochastic gradient descent on simulated data and demonstrated the existence of two learning phases.

We sought to validate their work on real-world datasets for a variety of network architectures, regularization strategies, and optimizers. Also, since training a model to perform well on unseen data is paramount, we examined ways in which to use the mutual information plane collected over training data to predict the model's relative performance on unseen test data.

Mutual Information

Formally, mutual information is the Kullback-Leibler divergence between a joint and its mean-field factorization. For the discrete case:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Mutual Information Plane as Applied to Deep Neural Networks

Theory:

Markov Chain: $Y \rightarrow X \rightarrow T^{(l)}$
DPI: $I(Y,X) \geq I(Y,T^{(l)})$
MI Plane: $(I(X,T^{(l)}), I(Y,T^{(l)}))$ in $\Re^2$

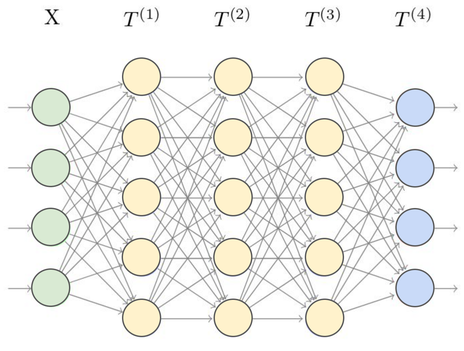
Task:

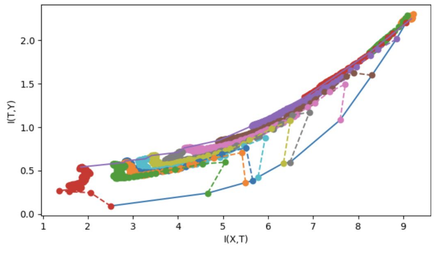
Classification: Cross-Entropy Loss

Data:

MNIST: 10 Outputs
CIFAR-10: 10 Outputs

Fully Connected Network:



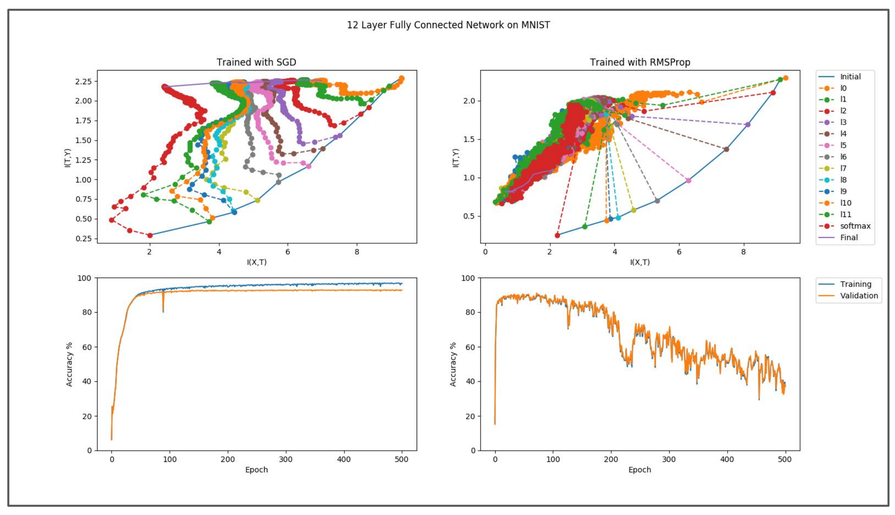


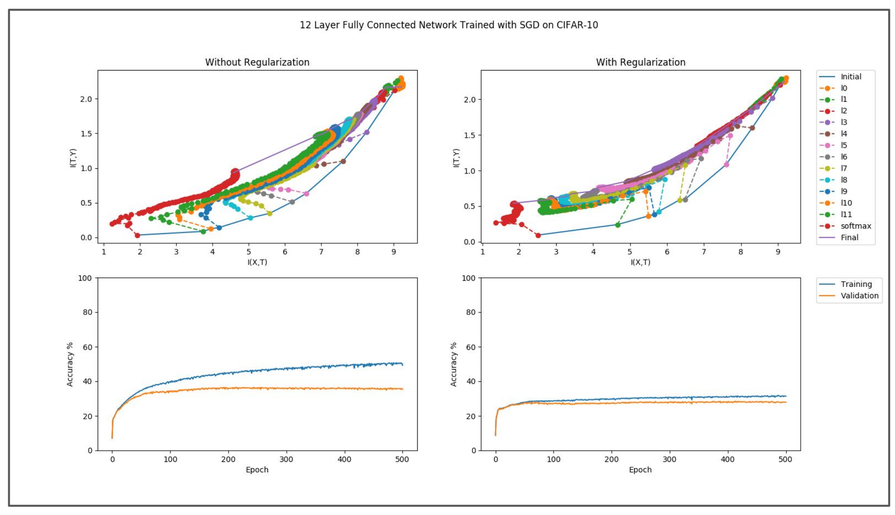
Learning Phases in Information Plane:

***Fitting Phase**:* Y-Axis Upward Movement

***Clustering Phase**:* X-Axis Leftward Movement

Experiments





Results

We used signals collected from the training set to predict which of two models would have better generalization performance on CIFAR-10. As shown, the training set's mutual information was the strongest predictor.

Network Type	Training Accuracy	Training Loss	Training M.I.
3 Layer FCN	36.7095%	60.5316%	**90.0603%**
6 Layer FCN	82.4646%	90.4204%	**91.5457%**
9 Layer FCN	85.0422%	87.4078%	**91.2167%**
12 Layer FCN	87.9248%	87.4568%	**91.7161%**
1 Layer CNN	97.3052%	95.6887%	**97.3693%**
2 Layer CNN	97.2769%	95.4170%	**97.3371%**

Future Work

Currently, the discrete treatment of continuous random variables prohibits the application to network layers with large output dimensions. In particular, this assumption results in unique values such that distributions collapse to the uniform distribution.

References

[1] R. Shwartz-Ziv and N. Tishby Opening the black box of deep neural networks via information. CoRR, abs/1703.00810, 2017.