Why does Unsupervised Pre-training Help deep learning?

Erhan, Dumitru, et al.

Overview

Deep Belief Network (DBN), Stacks of autoencoder
-> impressive result on vision and language datasets.

Unsupervised pre-training phase -> best results
-> **Why does unsupervised pre-training work so well?**

Possible explanations

1. Regularization (generalization)
2. Optimization

Experimental results

Better generalization
unsupervised pre-training gives substantially **lower test error** than no pre-training, for the same depth or for smaller on various vision datasets

Aid to optimization
the training error of the trained classifiers is low in all cases, with or without pre-training.
-> Such a result would make it difficult to distinguish between the optimization and regularization effects of pre-training.

constrain the **top layer to be small** (20 units instead of 500 and 1000).
-> the final **training errors are higher without pre-training.**

Experimental results

Distinct local minima
With 400 different random initializations, with or without pre-training,
each trajectory ends up in a different apparent local minimum

It is difficult to guarantee that these are indeed local minima but all tests performed
-> visual inspection of trajectories in function space, estimation of second derivatives in the directions of all the estimated eigenvectors of the Jacobian

The regions in function spaced reached without pre-training and with pre-training seem completely **disjoint**

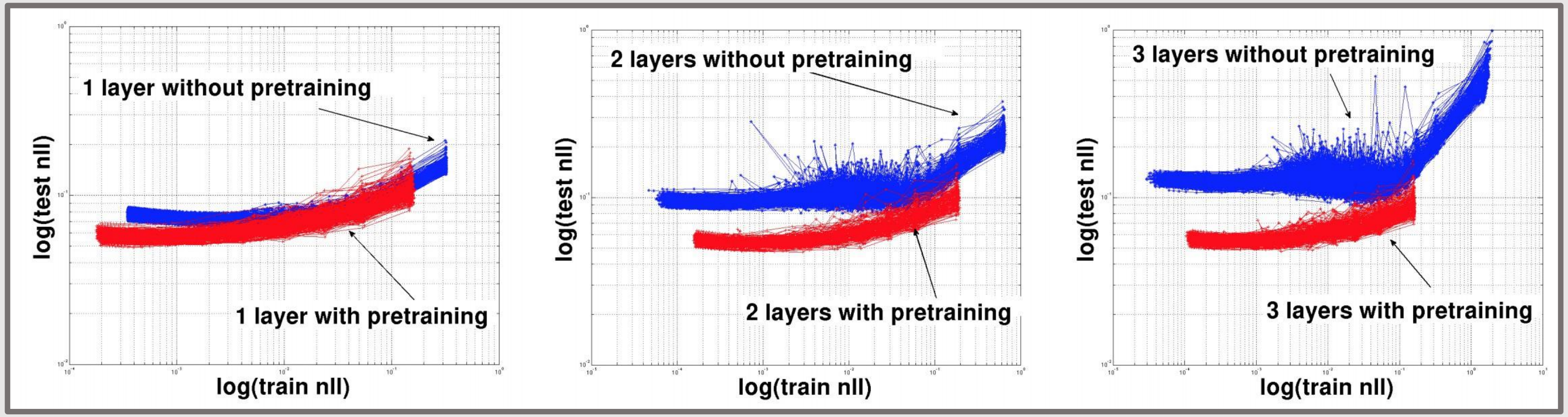
Lower variance
the variance of final test is larger without pre-training
-> regularization explanation, but does not exclude an optimization hypothesis either.

Hypothesis

1. Better optimization
Unsupervised pretraining puts the network in a region of parameter space where basins of attraction run deeper than when starting with random parameters. In simple words, the network starts near a global minimum.
In contrast to a local minimum, a global minimum means a lower **training error**.

2. Better regularization
Unsupervised pretraining puts the network in a region of parameter space in which systematically yields better generalization (lower **test error**).

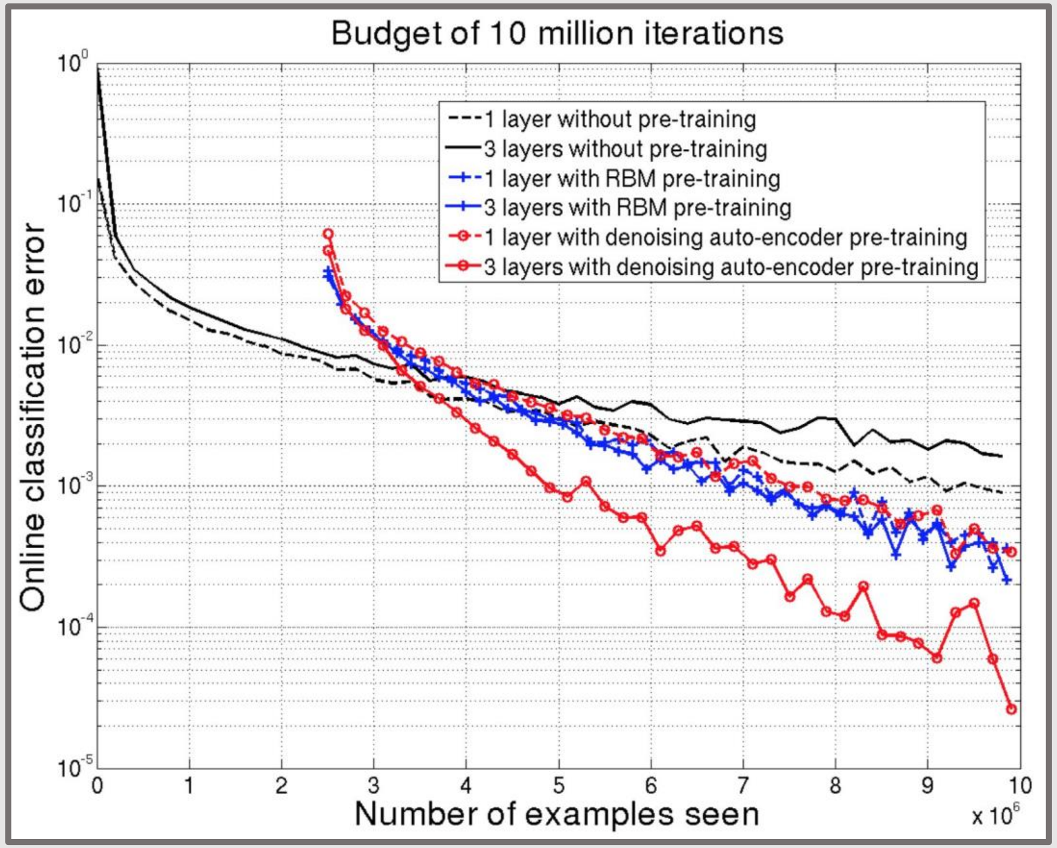
MNIST Experiments



the **test error is lower** with pre-training.
This **contradicts the better optimization** hypothesis because
it assumes pretraining would achieve lower training error

The pre-training regularizer is much better compared to L1/L2 regularizers.
-> L1/L2 regularizer : **decreases** as the data set grows
-> unsupervised pre-training : **maintained** as the data set grows.

MNIST Experiments

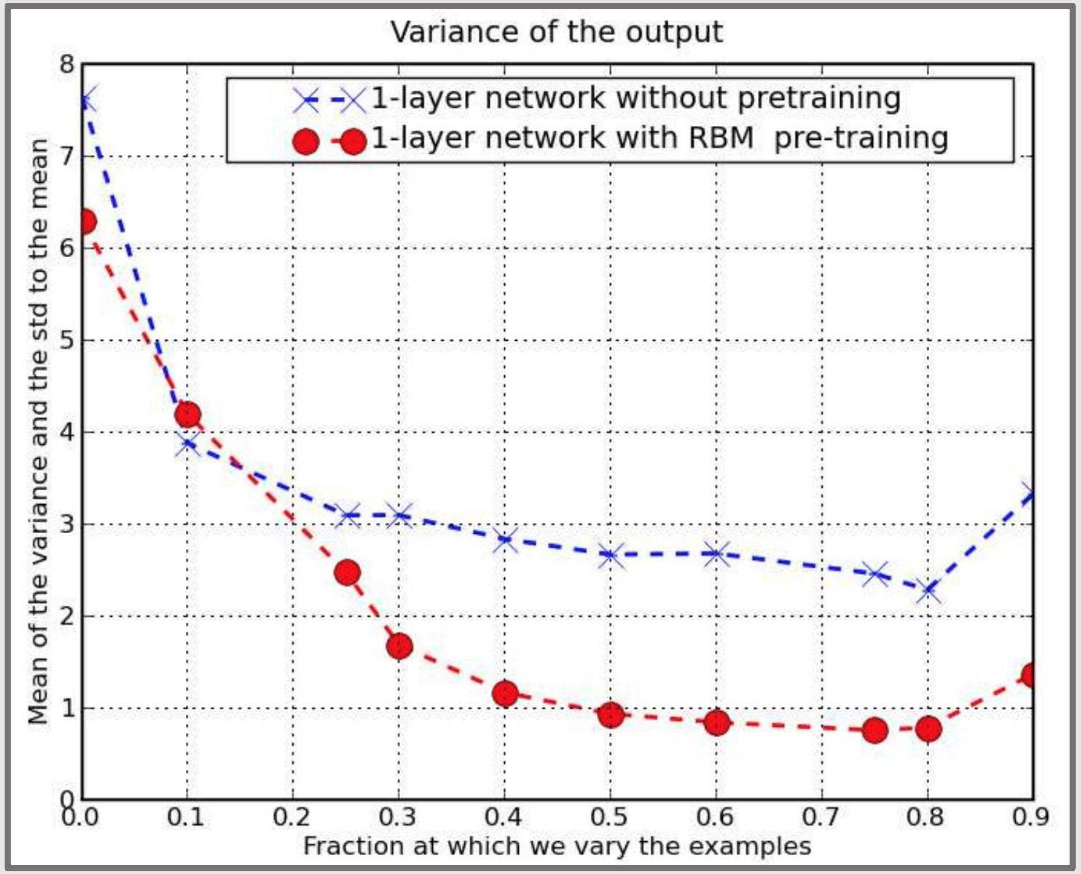


As the **dataset size** increases, the test error keeps decreasing with unsupervised pretraining.

MNIST Experiments

Quantify the impact of training samples' order on the network output variance.
-> High variance indicates that the order of the training samples significantly impacts the optimization problem.

-> **Early training samples** influence the output of the networks more than the ones at the end.
-> **this variance is lower for the pretrained** networks.
-> Finally, both networks (with and without pretraining) are **more influenced by the last examples used for optimization**, which is simply due to the fact that they use a stochastic gradient with a constant learning rate, where the most recent examples' gradient has a greater influence.



Analysis

ReLU, dropout, data augmentation, batch normalization (2010~)
-> Only supervised