Journal of Machine Learning Research 11 (2010) 625-660 Submitted 8/09; Published 2/10

Why Does Unsupervised Pre-training Help Deep Learning?

Dumitru Erhan[*] DUMITRU.ERHAN@UMONTREAL.CA
Yoshua Bengio YOSHUA.BENGIO@UMONTREAL.CA
Aaron Courville AARON.COURVILLE@UMONTREAL.CA
Pierre-Antoine Manzagol PIERRE-ANTOINE.MANZAGOL@UMONTREAL.CA
Pascal Vincent PASCAL.VINCENT@UMONTREAL.CA
Département d'informatique et de recherche opérationnelle
Université de Montréal
2920, chemin de la Tour
Montréal, Québec, H3T 1J8, Canada

Samy Bengio BENGIO@GOOGLE.COM
Google Research
1600 Amphitheatre Parkway
Mountain View, CA, 94043, USA

Editor: Léon Bottou

Abstract

Much recent research has been devoted to learning algorithms for deep architectures such as Deep Belief Networks and stacks of auto-encoder variants, with impressive results obtained in several areas, mostly on vision and language data sets. The best results obtained on supervised learning tasks involve an unsupervised learning component, usually in an unsupervised pre-training phase. Even though these new algorithms have enabled training deep models, many questions remain as to the nature of this difficult learning problem. The main question investigated here is the following: how does unsupervised pre-training work? Answering this questions is important if learning in deep architectures is to be further improved. We propose several explanatory hypotheses and test them through extensive simulations. We empirically show the influence of pre-training with respect to architecture depth, model capacity, and number of training examples. The experiments confirm and clarify the advantage of unsupervised pre-training. The results suggest that unsupervised pre-training guides the learning towards basins of attraction of minima that support better generalization from the training data set; the evidence from these results supports a regularization explanation for the effect of pre-training.

Keywords: deep architectures, unsupervised pre-training, deep belief networks, stacked denoising auto-encoders, non-convex optimization

1. Introduction

Deep learning methods aim at learning feature hierarchies with features from higher levels of the hierarchy formed by the composition of lower level features. They include learning methods for a wide array of *deep architectures* (Bengio, 2009 provides a survey), including neural networks with many hidden layers (Bengio et al., 2007; Ranzato et al., 2007; Vincent et al., 2008; Collobert and Weston, 2008) and graphical models with many levels of hidden variables (Hinton et al., 2006),

∗. Part of this work was done while Dumitru Erhan was at Google Research.

among others (Zhu et al., 2009; Weston et al., 2008). Theoretical results (Yao, 1985; Håstad, 1986; Håstad and Goldmann, 1991; Bengio et al., 2006), reviewed and discussed by Bengio and LeCun (2007), suggest that in order to learn the kind of complicated functions that can represent high-level abstractions (e.g., in vision, language, and other AI-level tasks), one may need *deep architectures*. The recent surge in experimental work in the field seems to support this notion, accumulating evidence that in challenging AI-related tasks—such as computer vision (Bengio et al., 2007; Ranzato et al., 2007; Larochelle et al., 2007; Ranzato et al., 2008; Lee et al., 2009; Mobahi et al., 2009; Osindero and Hinton, 2008), natural language processing (NLP) (Collobert and Weston, 2008; Weston et al., 2008), robotics (Hadsell et al., 2008), or information retrieval (Salakhutdinov and Hinton, 2007; Salakhutdinov et al., 2007)—deep learning methods significantly out-perform comparable but shallow competitors, and often match or beat the state-of-the-art.

These recent demonstrations of the potential of deep learning algorithms were achieved despite the serious challenge of training models with many layers of adaptive parameters. In virtually all instances of deep learning, the objective function is a highly non-convex function of the parameters, with the potential for many distinct *local minima* in the model parameter space. The principal difficulty is that not all of these minima provide equivalent generalization errors and, we suggest, that for deep architectures, the standard training schemes (based on random initialization) tend to place the parameters in regions of the parameters space that generalize poorly—as was frequently observed empirically but rarely reported (Bengio and LeCun, 2007).

The breakthrough to effective training strategies for deep architectures came in 2006 with the algorithms for training deep belief networks (DBN) (Hinton et al., 2006) and stacked auto-encoders (Ranzato et al., 2007; Bengio et al., 2007), which are all based on a similar approach: greedy layer-wise unsupervised pre-training followed by supervised fine-tuning. Each layer is pre-trained with an unsupervised learning algorithm, learning a nonlinear transformation of its input (the output of the previous layer) that captures the main variations in its input. This unsupervised pre-training sets the stage for a final training phase where the deep architecture is fine-tuned with respect to a supervised training criterion with gradient-based optimization. While the improvement in performance of trained deep models offered by the pre-training strategy is impressive, little is understood about the mechanisms underlying this success.

The objective of this paper is to explore, through extensive experimentation, how unsupervised pre-training works to render learning deep architectures more effective and why they appear to work so much better than traditional neural network training methods. There are a few reasonable hypotheses why unsupervised pre-training might work. One possibility is that unsupervised pre-training acts as a kind of network pre-conditioner, putting the parameter values in the appropriate range for further supervised training. Another possibility, suggested by Bengio et al. (2007), is that unsupervised pre-training initializes the model to a point in parameter space that somehow renders the optimization process more effective, in the sense of achieving a lower minimum of the empirical cost function.

Here, we argue that our experiments support a view of unsupervised pre-training as an unusual form of *regularization*: minimizing variance and introducing bias towards configurations of the parameter space that are useful for unsupervised learning. This perspective places unsupervised pre-training well within the family of recently developed semi-supervised methods. The unsupervised pre-training approach is, however, unique among semi-supervised training strategies in that it acts by defining a particular initialization point for standard supervised training rather than either modifying the supervised objective function (Barron, 1991) or explicitly imposing constraints on the parameters throughout training (Lasserre et al., 2006). This type of initialization-as-regularization strategy has precedence in the neural networks literature, in the shape of the early stopping idea (Sjöberg and Ljung, 1995; Amari et al., 1997), and in the Hidden Markov Models (HMM) community (Bahl et al., 1986; Povey and Woodland, 2002) where it was found that first training an HMM as a generative model was essential (as an initialization step) before fine-tuning it discriminatively. We suggest that, in the highly non-convex situation of training a deep architecture, defining a particular initialization point *implicitly* imposes constraints on the parameters in that it specifies which minima (out of a very large number of possible minima) of the cost function are allowed. In this way, it may be possible to think of unsupervised pre-training as being related to the approach of Lasserre et al. (2006).

Another important and distinct property of the unsupervised pre-training strategy is that in the standard situation of training using stochastic gradient descent, the beneficial generalization effects due to pre-training do not appear to diminish as the number of labeled examples grows very large. We argue that this is a consequence of the combination of the non-convexity (multi-modality) of the objective function and the dependency of the stochastic gradient descent method on example ordering. We find that early changes in the parameters have a greater impact on the final region (basin of attraction of the descent procedure) in which the learner ends up. In particular, unsupervised pre-training sets the parameter in a region from which better basins of attraction can be reached, in terms of generalization. Hence, although unsupervised pre-training is a regularizer, it can have a positive effect on the training objective when the number of training examples is large.

As previously stated, this paper is concerned with an experimental assessment of the various competing hypotheses regarding the role of unsupervised pre-training in the recent success of deep learning methods. To this end, we present a series of experiments design to pit these hypotheses against one another in an attempt to resolve some of the mystery surrounding the effectiveness of unsupervised pre-training.

In the first set of experiments (in Section 6), we establish the effect of unsupervised pre-training on improving the generalization error of trained deep architectures. In this section we also exploit dimensionality reduction techniques to illustrate how unsupervised pre-training affects the location of minima in parameter space.

In the second set of experiments (in Section 7), we directly compare the two alternative hypotheses (pre-training as a pre-conditioner; and pre-training as an optimization scheme) against the hypothesis that unsupervised pre-training is a regularization strategy. In the final set of experiments, (in Section 8), we explore the role of unsupervised pre-training in the online learning setting, where the number of available training examples grows very large. In these experiments, we test key aspects of our hypothesis relating to the topology of the cost function and the role of unsupervised pre-training in manipulating the region of parameter space from which supervised training is initiated.

Before delving into the experiments, we begin with a more in-depth view of the challenges in training deep architectures and how we believe unsupervised pre-training works towards overcoming these challenges.

2. The Challenges of Deep Learning

In this section, we present a perspective on why standard training of deep models through gradient backpropagation appears to be so difficult. First, it is important to establish what we mean in stating that training is difficult.

We believe the central challenge in training deep architectures is dealing with the strong dependencies that exist during training between the parameters across layers. One way to conceive the difficulty of the problem is that we must simultaneously:

1. adapt the lower layers in order to provide adequate input to the final (end of training) setting of the upper layers

2. adapt the upper layers to make good use of the final (end of training) setting of the lower layers.

The second problem is easy on its own (i.e., when the final setting of the other layers is known). It is not clear how difficult is the first one, and we conjecture that a particular difficulty arises when both sets of layers must be learned jointly, as the gradient of the objective function is limited to a local measure given the current setting of other parameters. Furthermore, because with enough capacity the top two layers can easily overfit the training set, training error does not necessarily reveal the difficulty in optimizing the lower layers. As shown in our experiments here, the standard training schemes tend to place the parameters in regions of the parameters space that generalize poorly.

A separate but related issue appears if we focus our consideration of traditional training methods for deep architectures on stochastic gradient descent. A sequence of examples along with an online gradient descent procedure defines a trajectory in parameter space, which converges in some sense (the error does not improve anymore, maybe because we are near a local minimum). The hypothesis is that small perturbations of that trajectory (either by initialization or by changes in which examples are seen when) have more effect early on. Early in the process of following the stochastic gradient, changes in the weights tend to increase their magnitude and, consequently, the amount of non-linearity of the network increases. As this happens, the set of regions accessible by stochastic gradient descent on samples of the training distribution becomes smaller. Early on in training small perturbations allow the model parameters to switch from one basin to a nearby one, whereas later on (typically with larger parameter values), it is unlikely to "escape" from such a basin of attraction. Hence the early examples can have a larger influence and, in practice, trap the model parameters in particular regions of parameter space that correspond to the specific and arbitrary ordering of the training examples.[1] An important consequence of this phenomenon is that even in the presence of a very large (effectively infinite) amounts of supervised data, stochastic gradient descent is subject to a degree of *overfitting* to the training data presented early in the training process. In that sense, unsupervised pre-training interacts intimately with the optimization process, and when the number of training examples becomes large, its positive effect is seen not only on generalization error but also on training error.

1. This process seems similar to the "critical period" phenomena observed in neuroscience and psychology (Bornstein, 1987).

3. Unsupervised Pre-training Acts as a Regularizer

As stated in the introduction, we believe that greedy layer-wise unsupervised pre-training overcomes the challenges of deep learning by introducing a useful prior to the *supervised fine-tuning* training procedure. We claim that the regularization effect is a consequence of the pre-training procedure establishing an initialization point of the fine-tuning procedure inside a region of parameter space in which the parameters are henceforth restricted. The parameters are restricted to a relatively small volume of parameter space that is delineated by the boundary of the *local basin of attraction* of the supervised fine-tuning cost function.

The pre-training procedure increases the magnitude of the weights and in standard deep models, with a sigmoidal nonlinearity, this has the effect of rendering both the function more nonlinear and the cost function locally more complicated with more topological features such as peaks, troughs and plateaus. The existence of these topological features renders the parameter space locally more difficult to travel significant distances via a gradient descent procedure. This is the core of the restrictive property imposed by the pre-training procedure and hence the basis of its regularizing properties.

But unsupervised pre-training restricts the parameters to particular regions: those that correspond to capturing structure in the input distribution $P(X)$. To simply state that unsupervised pre-training is a regularization strategy somewhat undermines the significance of its effectiveness. Not all regularizers are created equal and, in comparison to standard regularization schemes such as L_1 and L_2 parameter penalization, unsupervised pre-training is dramatically effective. We believe the credit for its success can be attributed to the unsupervised training criteria optimized during unsupervised pre-training.

During each phase of the greedy unsupervised training strategy, layers are trained to represent the dominant factors of variation extant in the data. This has the effect of leveraging knowledge of X to form, at each layer, a representation of X consisting of statistically reliable features of X that can then be used to predict the output (usually a class label) Y. This perspective places unsupervised pre-training well within the family of learning strategies collectively know as semi-supervised methods. As with other recent work demonstrating the effectiveness of semi-supervised methods in regularizing model parameters, we claim that the effectiveness of the unsupervised pre-training strategy is limited to the extent that learning $P(X)$ is helpful in learning $P(Y|X)$. Here, we find transformations of X—learned features—that are predictive of the main factors of variation in $P(X)$, and when the pre-training strategy is effective,[2] some of these learned features of X are also predictive of Y. In the context of deep learning, the greedy unsupervised strategy may also have a special function. To some degree it resolves the problem of simultaneously learning the parameters at all layers (mentioned in Section 2) by introducing a proxy criterion. This proxy criterion encourages significant factors of variation, present in the input data, to be represented in intermediate layers.

To clarify this line of reasoning, we can formalize the effect of unsupervised pre-training in inducing a prior distribution over the parameters. Let us assume that parameters are forced to be chosen in a bounded region $\mathcal{S} \subset \mathbb{R}^d$. Let $\mathcal{S}$ be split in regions $\{R_k\}$ that are the basins of attraction of descent procedures in the training error (note that $\{R_k\}$ depends on the training set, but the dependency decreases as the number of examples increases). We have $\cup_k R_k = \mathcal{S}$ and $R_i \cap R_j = \emptyset$ for $i \neq j$. Let $v_k = \int 1_{\theta \in R_k} d\theta$ be the volume associated with region R_k (where θ are our model's

2. Acting as a form of (data-dependent) "prior" on the parameters, as we are about to formalize.

parameters). Let r_k be the probability that a purely random initialization (according to our initialization procedure, which factorizes across parameters) lands in R_k, and let π_k be the probability that pre-training (following a random initialization) lands in R_k, that is, $\sum_k r_k = \sum_k \pi_k = 1$. We can now take into account the initialization procedure as a regularization term:

$$\text{regularizer} = -\log P(\theta).$$

For pre-trained models, the prior is

$$P_{\text{pre-training}}(\theta) = \sum_k 1_{\theta \in R_k} \pi_k / v_k.$$

For the models without unsupervised pre-training, the prior is

$$P_{\text{no-pre-training}}(\theta) = \sum_k 1_{\theta \in R_k} r_k / v_k.$$

One can verify that $P_{\text{pre-training}}(\theta \in R_k) = \pi_k$ and $P_{\text{no-pre-training}}(\theta \in R_k) = r_k$. When π_k is tiny, the penalty is high when $\theta \in R_k$, with unsupervised pre-training. The derivative of this regularizer is zero almost everywhere because we have chosen a uniform prior inside each region R_k. Hence, to take the regularizer into account, and having a generative model $P_{\text{pre-training}}(\theta)$ for θ (i.e., this is the unsupervised pre-training procedure), it is reasonable to sample an initial θ from it (knowing that from this point on the penalty will not increase during the iterative minimization of the training criterion), and this is exactly how the pre-trained models are obtained in our experiments.

Note that this formalization is just an illustration: it is there to simply show how one could conceptually think of an initialization point as a regularizer and should not be taken as a literal interpretation of how regularization is explicitly achieved, since we do not have an analytic formula for computing the π_k's and v_k's. Instead these are implicitly defined by the whole unsupervised pre-training procedure.

4. Previous Relevant Work

We start with an overview of the literature on semi-supervised learning (SSL), since the SSL framework is essentially the one in which we operate as well.

4.1 Related Semi-Supervised Methods

It has been recognized for some time that generative models are less prone to overfitting than discriminant ones (Ng and Jordan, 2002). Consider input variable X and target variable Y. Whereas a discriminant model focuses on $P(Y|X)$, a generative model focuses on $P(X,Y)$ (often parametrized as $P(X|Y)P(Y)$), that is, it also cares about getting $P(X)$ right, which can reduce the freedom of fitting the data when the ultimate goal is only to predict Y given X.

Exploiting information about $P(X)$ to improve generalization of a classifier has been the driving idea behind semi-supervised learning (Chapelle et al., 2006). For example, one can use unsupervised learning to map X into a representation (also called embedding) such that two examples $\mathbf{x}_1$ and $\mathbf{x}_2$ that belong to the same cluster (or are reachable through a short path going through neighboring examples in the training set) end up having nearby embeddings. One can then use supervised learning (e.g., a linear classifier) in that new space and achieve better generalization in many cases (Belkin and Niyogi, 2002; Chapelle et al., 2003). A long-standing variant of this approach is the application of Principal Components Analysis as a pre-processing step before applying a classifier (on the projected data). In these models the data is first transformed in a new representation using unsupervised learning, and a supervised classifier is stacked on top, learning to map the data in this new representation into class predictions.

Instead of having separate unsupervised and supervised components in the model, one can consider models in which $P(X)$ (or $P(X,Y)$) and $P(Y|X)$ share parameters (or whose parameters are connected in some way), and one can trade-off the supervised criterion $-\log P(Y|X)$ with the unsupervised or generative one ($-\log P(X)$ or $-\log P(X,Y)$). It can then be seen that the generative criterion corresponds to a particular form of prior (Lasserre et al., 2006), namely that the structure of $P(X)$ is connected to the structure of $P(Y|X)$ in a way that is captured by the shared parametrization. By controlling how much of the generative criterion is included in the total criterion, one can find a better trade-off than with a purely generative or a purely discriminative training criterion (Lasserre et al., 2006; Larochelle and Bengio, 2008).

In the context of deep architectures, a very interesting application of these ideas involves adding an unsupervised embedding criterion at each layer (or only one intermediate layer) to a traditional supervised criterion (Weston et al., 2008). This has been shown to be a powerful semi-supervised learning strategy, and is an alternative to the kind of algorithms described and evaluated in this paper, which also combine unsupervised learning with supervised learning.

In the context of scarcity of labelled data (and abundance of unlabelled data), deep architectures have shown promise as well. Salakhutdinov and Hinton (2008) describe a method for learning the covariance matrix of a Gaussian Process, in which the usage of unlabelled examples for modeling $P(X)$ improves $P(Y|X)$ quite significantly. Note that such a result is to be expected: with few labelled samples, modeling $P(X)$ usually helps. Our results show that even in the context of *abundant labelled data*, unsupervised pre-training still has a pronounced positive effect on generalization: a somewhat surprising conclusion.

4.2 Early Stopping as a Form of Regularization

We stated that pre-training as initialization can be seen as restricting the optimization procedure to a relatively small volume of parameter space that corresponds to a local basin of attraction of the supervised cost function. Early stopping can be seen as having a similar effect, by constraining the optimization procedure to a region of the parameter space that is close to the initial configuration of parameters. With τ the number of training iterations and η the learning rate used in the update procedure, $\tau\eta$ can be seen as the reciprocal of a regularization parameter. Indeed, restricting either quantity restricts the area of parameter space reachable from the starting point. In the case of the optimization of a simple linear model (initialized at the origin) using a quadratic error function and simple gradient descent, early stopping will have a similar effect to traditional regularization.

Thus, in both pre-training and early stopping, the parameters of the supervised cost function are constrained to be close to their initial values.[3] A more formal treatment of early stopping as regularization is given by Sjöberg and Ljung (1995) and Amari et al. (1997). There is no equivalent treatment of pre-training, but this paper sheds some light on the effects of such initialization in the case of deep architectures.

3. In the case of pre-training the "initial values" of the parameters for the supervised phase are those that were obtained at the end of pre-training.

5. Experimental Setup and Methodology

In this section, we describe the setting in which we test the hypothesis introduced in Section 3 and previously proposed hypotheses. The section includes a description of the deep architectures used, the data sets and the details necessary to reproduce our results.

5.1 Models

All of the successful methods (Hinton et al., 2006; Hinton and Salakhutdinov, 2006; Bengio et al., 2007; Ranzato et al., 2007; Vincent et al., 2008; Weston et al., 2008; Ranzato et al., 2008; Lee et al., 2008) in the literature for training deep architectures have something in common: they rely on an unsupervised learning algorithm that provides a training signal at the level of a single layer. Most work in two main phases. In a first phase, *unsupervised pre-training*, all layers are initialized using this layer-wise unsupervised learning signal. In a second phase, *fine-tuning*, a global training criterion (a prediction error, using labels in the case of a supervised task) is minimized. In the algorithms initially proposed (Hinton et al., 2006; Bengio et al., 2007; Ranzato et al., 2007), the unsupervised pre-training is done in a greedy layer-wise fashion: at stage k, the k-th layer is trained (with respect to an unsupervised criterion) using as input the output of the previous layer, and while the previous layers are kept fixed.

We shall consider two deep architectures as representatives of two families of models encountered in the deep learning literature.

5.1.1 DEEP BELIEF NETWORKS

The first model is the Deep Belief Net (DBN) by Hinton et al. (2006), obtained by training and stacking several layers of Restricted Boltzmann Machines (RBM) in a greedy manner. Once this stack of RBMs is trained, it can be used to initialize a multi-layer neural network for classification.

An RBM with n hidden units is a Markov Random Field (MRF) for the joint distribution between hidden variables h_i and observed variables x_j such that $P(\mathbf{h}|\mathbf{x})$ and $P(\mathbf{x}|\mathbf{h})$ factorize, that is, $P(\mathbf{h}|\mathbf{x}) = \prod_i P(h_i|\mathbf{x})$ and $P(\mathbf{x}|\mathbf{h}) = \prod_j P(x_j|\mathbf{h})$. The sufficient statistics of the MRF are typically h_i, x_j and $h_i x_j$, which gives rise to the following joint distribution:

$$P(\mathbf{x},\mathbf{h}) \propto e^{\mathbf{h}'W\mathbf{x}+b'\mathbf{x}+c'\mathbf{h}}$$

with corresponding parameters $\theta = (W, b, c)$ (with $'$ denoting transpose, c_i associated with h_i, b_j with x_j, and W_{ij} with $h_i x_j$). If we restrict h_i and x_j to be binary units, it is straightforward to show that

$$\begin{aligned} P(\mathbf{x}|\mathbf{h}) &= \prod_j P(x_j|\mathbf{h}) \quad \text{with} \\ P(x_j = 1|\mathbf{h}) &= \text{sigmoid}(b_j + \sum_i W_{ij} h_i). \end{aligned}$$

where $\text{sigmoid}(\mathbf{a}) = 1/(1+\exp(-\mathbf{a}))$ (applied element-wise on a vector $\mathbf{a}$), and $P(\mathbf{h}|\mathbf{x})$ also has a similar form:

$$\begin{aligned} P(\mathbf{h}|\mathbf{x}) &= \prod_i P(h_i|\mathbf{x}) \quad \text{with} \\ P(h_i = 1|\mathbf{x}) &= \text{sigmoid}(c_i + \sum_j W_{ij} x_j). \end{aligned}$$

The RBM form can be generalized to other conditional distributions besides the binomial, including continuous variables. Welling et al. (2005) describe a generalization of RBM models to conditional distributions from the exponential family.

RBM models can be trained by approximate stochastic gradient descent. Although $P(\mathbf{x})$ is not tractable in an RBM, the Contrastive Divergence estimator (Hinton, 2002) is a good stochastic approximation of $\frac{\partial \log P(\mathbf{x})}{\partial \theta}$, in that it very often has the same sign (Bengio and Delalleau, 2009).

A DBN is a multi-layer generative model with layer variables h_0 (the input or visible layer), h_1, h_2, etc. The top two layers have a joint distribution which is an RBM, and $P(h_k|h_{k+1})$ are parametrized in the same way as for an RBM. Hence a 2-layer DBN is an RBM, and a stack of RBMs share parametrization with a corresponding DBN. The contrastive divergence update direction can be used to initialize each layer of a DBN as an RBM, as follows. Consider the first layer of the DBN trained as an RBM P_1 with hidden layer h_1 and visible layer v_1. We can train a second RBM P_2 that models (in its visible layer) the samples h_1 from $P_1(h_1|v_1)$ when v_1 is sampled from the training data set. It can be shown that this maximizes a lower bound on the log-likelihood of the DBN. The number of layers can be increased greedily, with the newly added top layer trained as an RBM to model the samples produced by chaining the posteriors $P(h_k|h_{k-1})$ of the lower layers (starting from h_0 from the training data set).

The parameters of a DBN or of a stack of RBMs also correspond to the parameters of a deterministic feed-forward multi-layer neural network. The i-th unit of the k-th layer of the neural network outputs $\hat{h}_{ki} = \text{sigmoid}(c_{ki} + \sum_j W_{kij}\hat{h}_{k-1,j})$, using the parameters c_k and W_k of the k-th layer of the DBN. Hence, once the stack of RBMs or the DBN is trained, one can use those parameters to initialize the first layers of a corresponding multi-layer neural network. One or more additional layers can be added to map the top-level features $\hat{h}_k$ to the predictions associated with a target variable (here the probabilities associated with each class in a classification task). Bengio (2009) provides more details on RBMs and DBNs, and a survey of related models and deep architectures.

5.1.2 STACKED DENOISING AUTO-ENCODERS

The second model, by Vincent et al. (2008), is the so-called Stacked Denoising Auto-Encoder (SDAE). It borrows the greedy principle from DBNs, but uses denoising auto-encoders as a building block for unsupervised modeling. An auto-encoder learns an encoder $h(\cdot)$ and a decoder $g(\cdot)$ whose composition approaches the identity for examples in the training set, that is, $g(h(\mathbf{x})) \approx \mathbf{x}$ for $\mathbf{x}$ in the training set.

Assuming that some constraint prevents $g(h(\cdot))$ from being the identity for arbitrary arguments, the auto-encoder has to capture statistical structure in the training set in order to minimize reconstruction error. However, with a high capacity code ($h(\mathbf{x})$ has too many dimensions), a regular auto-encoder could potentially learn a trivial encoding. Note that there is an intimate connection between minimizing reconstruction error for auto-encoders and contrastive divergence training for RBMs, as both can be shown to approximate a log-likelihood gradient (Bengio and Delalleau, 2009).

The *denoising auto-encoder* (Vincent et al., 2008; Seung, 1998; LeCun, 1987; Gallinari et al., 1987) is a stochastic variant of the ordinary auto-encoder with the distinctive property that even with a high capacity model, it cannot learn the identity mapping. A denoising autoencoder is explicitly trained to denoise a corrupted version of its input. Its training criterion can also be viewed as a variational lower bound on the likelihood of a specific generative model. It has been shown on an array of data sets to perform significantly better than ordinary auto-encoders and similarly or better than RBMs when stacked into a deep supervised architecture (Vincent et al., 2008). Another way to prevent regular auto-encoders with more code units than inputs to learn the identity is to restrict the capacity of the representation by imposing sparsity on the code (Ranzato et al., 2007, 2008).

We now summarize the training algorithm of the Stacked Denoising Auto-Encoders. More details are given by Vincent et al. (2008). Each denoising auto-encoder operates on its inputs $\mathbf{x}$, either the raw inputs or the outputs of the previous layer. The denoising auto-encoder is trained to reconstruct $\mathbf{x}$ from a stochastically corrupted (noisy) transformation of it. The output of each denoising auto-encoder is the "code vector" $h(\mathbf{x})$, not to confuse with the reconstruction obtained by applying the decoder to that code vector. In our experiments $h(\mathbf{x}) = \text{sigmoid}(\mathbf{b} + W\mathbf{x})$ is an ordinary neural network layer, with hidden unit biases $\mathbf{b}$, and weight matrix W. Let $C(\mathbf{x})$ represent a stochastic corruption of $\mathbf{x}$. As done by Vincent et al. (2008), we set $C_i(\mathbf{x}) = x_i$ or 0, with a random subset (of a fixed size) selected for zeroing. We have also considered a salt and pepper noise, where we select a random subset of a fixed size and set $C_i(\mathbf{x}) = \text{Bernoulli}(0.5)$. The denoised "reconstruction" is obtained from the noisy input with $\hat{\mathbf{x}} = \text{sigmoid}(\mathbf{c} + W^T h(C(\mathbf{x})))$, using biases $\mathbf{c}$ and the transpose of the feed-forward weights W. In the experiments on images, both the raw input x_i and its reconstruction $\hat{x}_i$ for a particular pixel i can be interpreted as a Bernoulli probability for that pixel: the probability of painting the pixel as black at that location. We denote $\text{CE}(\mathbf{x}||\hat{\mathbf{x}}) = \sum_i \text{CE}(x_i||\hat{x}_i)$ the sum of the component-wise cross-entropy between the Bernoulli probability distributions associated with each element of $\mathbf{x}$ and its reconstruction probabilities $\hat{\mathbf{x}}$: $\text{CE}(\mathbf{x}||\hat{\mathbf{x}}) = -\sum_i (x_i \log \hat{x}_i + (1 - x_i)\log(1 - \hat{x}_i))$. The Bernoulli model only makes sense when the input components and their reconstruction are in $[0,1]$; another option is to use a Gaussian model, which corresponds to a Mean Squared Error (MSE) criterion.

With either DBN or SDAE, an output logistic regression layer is added after unsupervised training. This layer uses softmax (multinomial logistic regression) units to estimate $P(\text{class}|\mathbf{x}) = \text{softmax}_{\text{class}}(\mathbf{a})$, where a_i is a linear combination of outputs from the top hidden layer. The whole network is then trained as usual for multi-layer perceptrons, to minimize the output (negative log-likelihood) prediction error.

5.2 Data Sets

We experimented on three data sets, with the motivation that our experiments would help understand previously presented results with deep architectures, which were mostly with the MNIST data set and variations (Hinton et al., 2006; Bengio et al., 2007; Ranzato et al., 2007; Larochelle et al., 2007; Vincent et al., 2008):

`MNIST` the digit classification data set by LeCun et al. (1998), containing 60,000 training and 10,000 testing examples of 28x28 handwritten digits in gray-scale.

`InfiniteMNIST` a data set by Loosli et al. (2007), which is an extension of `MNIST` from which one can obtain a quasi-infinite number of examples. The samples are obtained by performing random elastic deformations of the original `MNIST` digits. In this data set, there is only one set of examples, and the models will be compared by their (online) performance on it.

`Shapeset` is a synthetic data set with a controlled range of geometric invariances. The underlying task is binary classification of 10×10 images of triangles and squares. The examples show images of shapes with many variations, such as size, orientation and gray-level. The data set is composed of 50000 training, 10000 validation and 10000 test images.[4]

5.3 Setup

The models used are

1. Deep Belief Networks containing Bernoulli RBM layers,
2. Stacked Denoising Auto-Encoders with Bernoulli input units, and
3. standard feed-forward multi-layer neural networks,

each with 1–5 hidden layers. Each hidden layer contains the same number of hidden units, which is a hyperparameter. The other hyperparameters are the unsupervised and supervised learning rates, the L_2 penalty / weight decay,[5] and the fraction of stochastically corrupted inputs (for the SDAE). For `MNIST`, the number of supervised and unsupervised passes through the data (epochs) is 50 and 50 per layer, respectively. With `InfiniteMNIST`, we perform 2.5 million unsupervised updates followed by 7.5 million supervised updates.[6] The standard feed-forward networks are trained using 10 million supervised updates. For `MNIST`, model selection is done by choosing the hyperparameters that optimize the supervised (classification) error on the validation set. For `InfiniteMNIST`, we use the average online error over the last million examples for hyperparameter selection. In all cases, purely stochastic gradient updates are applied.

The experiments involve the training of deep architectures with a variable number of layers with and without unsupervised pre-training. For a given layer, weights are initialized using random samples from uniform$[-1/\sqrt{k}, 1/\sqrt{k}]$, where k is the number of connections that a unit receives from the previous layer (the fan-in). Either supervised gradient descent or unsupervised pre-training follows.

In most cases (for `MNIST`), we first launched a number of experiments using a cross-product of hyperparameter values[7] applied to 10 different random initialization seeds. We then selected the hyperparameter sets giving the best validation error for each combination of model (with or without pre-training), number of layers, and number of training iterations. Using these hyper-parameters, we launched experiments using an additional 400 initialization seeds. For `InfiniteMNIST`, only one seed is considered (an arbitrarily chosen value).

In the discussions below we sometimes use the word **apparent local minimum** to mean the solution obtained after training, when no further noticeable progress seems achievable by stochastic gradient descent. It is possible that these are not really near a true local minimum (there could be a tiny ravine towards significant improvement, not accessible by gradient descent), but it is clear that these end-points represent regions where gradient descent is stuck. Note also that when we write of number of layers it is to be understood as the number of *hidden* layers in the network.

4. The data set can be downloaded from `http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/ShapesetDataForJMLR`.
5. A penalizing term $\lambda||\theta||_2^2$ is added to the supervised objective, where θ are the weights of the network, and λ is a hyper-parameter modulating the strength of the penalty.
6. The number of examples was chosen to be as large as possible, while still allowing for the exploration a variety of hyper-parameters.
7. Number of hidden units $\in \{400, 800, 1200\}$; learning rate $\in \{0.1, 0.05, 0.02, 0.01, 0.005\}$; ℓ_2 penalty coefficient $\lambda \in \{10^{-4}, 10^{-5}, 10^{-6}, 0\}$; pre-training learning rate $\in \{0.01, 0.005, 0.002, 0.001, 0.0005\}$; corruption probability $\in \{0.0, 0.1, 0.25, 0.4\}$; tied weights $\in \{\text{yes}, \text{no}\}$.

6. The Effect of Unsupervised Pre-training

We start by a presentation of large-scale simulations that were intended to confirm some of the previously published results about deep architectures. In the process of analyzing them, we start making connections to our hypotheses and motivate the experiments that follow.

6.1 Better Generalization

When choosing the number of units per layer, the learning rate and the number of training iterations to optimize classification error on the validation set, unsupervised pre-training gives substantially lower test classification error than no pre-training, for the same depth or for smaller depth on various vision data sets (Ranzato et al., 2007; Bengio et al., 2007; Larochelle et al., 2009, 2007; Vincent et al., 2008) no larger than the `MNIST` digit data set (experiments reported from 10,000 to 50,000 training examples).

Such work was performed with only one or a handful of different random initialization seeds, so one of the goals of this study was to ascertain the effect of the random seed used when initializing ordinary neural networks (deep or shallow) and the pre-training procedure. For this purpose, between 50 and 400 different seeds were used to obtain the graphics on `MNIST`.

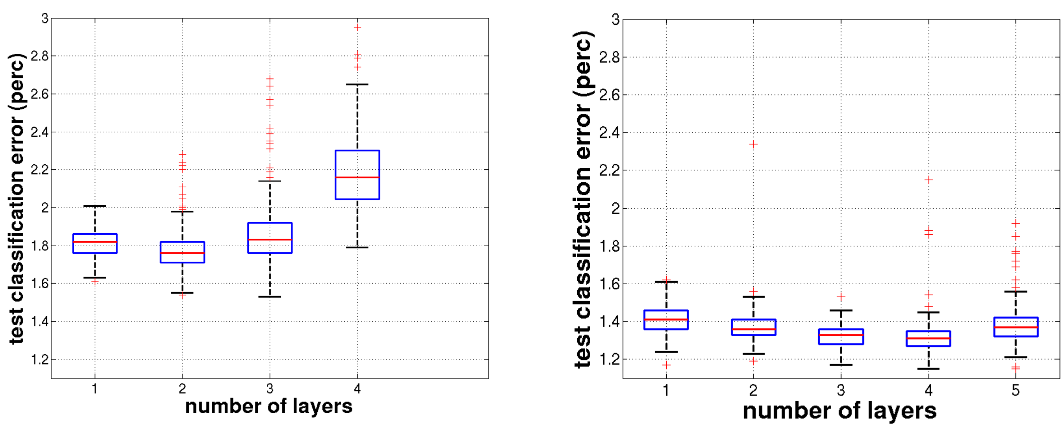

Figure 1: Effect of depth on performance for a model trained (**left**) without unsupervised pre-training and (**right**) with unsupervised pre-training, for 1 to 5 hidden layers (networks with 5 layers failed to converge to a solution, without the use of unsupervised pre-training). Experiments on `MNIST`. Box plots show the distribution of errors associated with 400 different initialization seeds (top and bottom quartiles in box, plus outliers beyond top and bottom quartiles). Other hyperparameters are optimized away (on the validation set). *Increasing depth seems to increase the probability of finding poor apparent local minima.*

Figure 1 shows the resulting distribution of test classification error, obtained with and without pre-training, as we increase the depth of the network. Figure 2 shows these distributions as histograms in the case of 1 and 4 layers. As can be seen in Figure 1, unsupervised pre-training allows classification error to go down steadily as we move from 1 to 4 hidden layers, whereas without pre-training the error goes up after 2 hidden layers. It should also be noted that we were unable to effectively train 5-layer models without use of unsupervised pre-training. Not only is the error obtained on average with unsupervised pre-training systematically lower than without the pre-training, it appears also more robust to the random initialization. With unsupervised pre-training the variance stays at about the same level up to 4 hidden layers, with the number of bad outliers growing slowly.

Contrast this with the case without pre-training: the variance and number of bad outliers grows sharply as we increase the number of layers beyond 2. The gain obtained with unsupervised pre-training is more pronounced as we increase the number of layers, as is the gain in robustness to random initialization. This can be seen in Figure 2. The increase in error variance and mean for deeper architectures without pre-training suggests that **increasing depth increases the probability of finding poor apparent local minima** when starting from random initialization. It is also interesting to note the low variance and small spread of errors obtained with 400 seeds with unsupervised pre-training: it suggests that **unsupervised pre-training is robust with respect to the random initialization seed** (the one used to initialize parameters before pre-training).

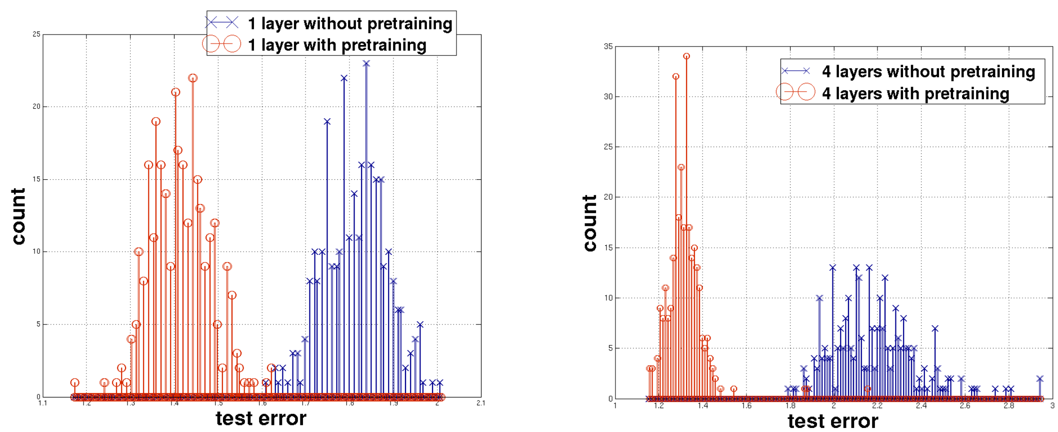

Figure 2: Histograms presenting the test errors obtained on MNIST using models trained with or without pre-training (400 different initializations each). **Left**: 1 hidden layer. **Right**: 4 hidden layers.

These experiments show that the variance of final test error with respect to initialization random seed is larger without pre-training, and this effect is magnified for deeper architectures. It should however be noted that there is a limit to the success of this technique: performance degrades for 5 layers on this problem.

6.2 Visualization of Features

Figure 3 shows the weights (called filters) of the first layer of the DBN before and after supervised fine-tuning. For visualizing what units do on the 2nd and 3rd layer, we used the activation maximization technique described by Erhan et al. (2009): to visualize what a unit responds most to, the method looks for the bounded input pattern that maximizes the activation of a given unit. This is an optimization problem which is solved by performing gradient ascent in the space of the inputs, to find a local maximum of the activation function. Interestingly, nearly the same maximal activation input pattern is recovered from most random initializations of the input pattern.

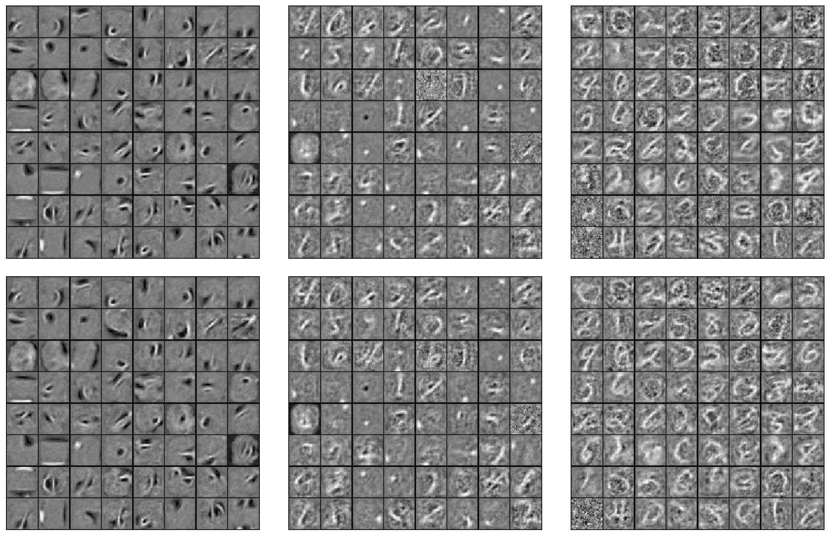

Figure 3: Visualization of filters learned by a DBN trained on `InfiniteMNIST`. The top figures contain a visualization of filters after pre-training, while the bottoms ones picture the same units after supervised fine-tuning; from left to right: units from the 1st, 2nd and 3rd layers, respectively.

For comparison, we have also visualized the filters of a network for 1–3 layers in which no pre-training was performed (Figure 4). While the first layer filters do seem to correspond to localized features, 2nd and 3rd layers are not as interpretable anymore. Qualitatively speaking, filters from the bottom row of Figure 3 and those from Figure 4 have little in common, which is an interesting conclusion in itself. In addition, there seems to be more interesting visual structures in the features learned in networks with unsupervised pre-training.

Several interesting conclusions can be drawn from Figure 3. First, supervised fine-tuning (after unsupervised pre-training), even with 7.5 million updates, does not change the weights in a significant way (at least visually): they seem stuck in a certain region of weight space, and the sign of weights does not change after fine-tuning (hence the same pattern is seen visually). Second, different layers change differently: the first layer changes least, while supervised training has more effect when performed on the 3rd layer. Such observations are consistent with the predictions made by our hypothesis: namely that the early dynamics of stochastic gradient descent, the dynamics induced by unsupervised pre-training, can "lock" the training in a region of the parameter space that is essentially inaccessible for models that are trained in a purely supervised way.

Finally, the features increase in complexity as we add more layers. First layer weights seem to encode basic stroke-like detectors, second layer weights seem to detect digit parts, while top layer weights detect entire digits. The features are more complicated as we add more layers, and displaying only one image for each "feature" does not do justice to the non-linear nature of that feature. For example, it does not show the *set of patterns* on which the feature is highly active (or highly inactive).

While Figures 3–4 show only the filters obtained on `InfiniteMNIST`, the visualizations are similar when applied on `MNIST`. Likewise, the features obtained with SDAE result in qualitatively similar conclusions; Erhan et al. (2009) gives more details.

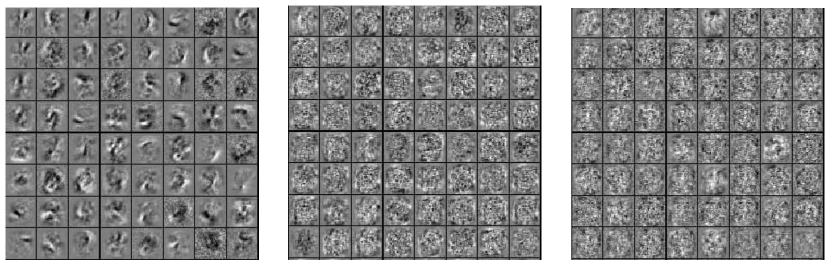

Figure 4: Visualization of filters learned by a network without pre-training, trained on `InfiniteMNIST`. The filters are shown after supervised training; from left to right: units from the 1st, 2nd and 3rd layers, respectively.

6.3 Visualization of Model Trajectories During Learning

Visualizing the learned features allows for a qualitative comparison of the training strategies for deep architectures. However it is not useful for investigating how these strategies are influenced by random initialization, as the features learned from multiple initializations look similar. If it was possible for us to visualize a variety of models at the same time, it would allow us to explore our hypothesis, and ascertain to what degree and how the set of pre-trained models (for different random seeds) is far from the set of models without pre-training. Do these two sets cover very different regions in parameter space? Are parameter trajectories getting stuck in many different apparent local minima?

Unfortunately, it is not possible to directly compare parameter values of two architectures, because many permutations of the same parameters give rise to the same model. However, one can take a functional approximation approach in which we compare the function (from input to output) represented by each network, rather than comparing the parameters. The function is the infinite ordered set of output values associated with all possible inputs, and it can be approximated with a finite number of inputs (preferably plausible ones). To visualize the trajectories followed during training, we use the following procedure. For a given model, we compute and concatenate all its outputs on the test set examples as one long vector summarizing where it stands in "function space". We get one such vector for each partially trained model (at each training iteration). This allows us to plot many learning trajectories, one for each initialization seed, with or without pre-training. Using a dimensionality reduction algorithm we then map these vectors to a two-dimensional space for visualization.[8] Figures 5 and 6 present the results using dimensionality reduction techniques that

8. Note that we can and do project the models with and without pre-training at the same time, so as to visualize them in the same space.

focus respectively on local[9] and global structure.[10] Each point is colored according to the training iteration, to help follow the trajectory movement.

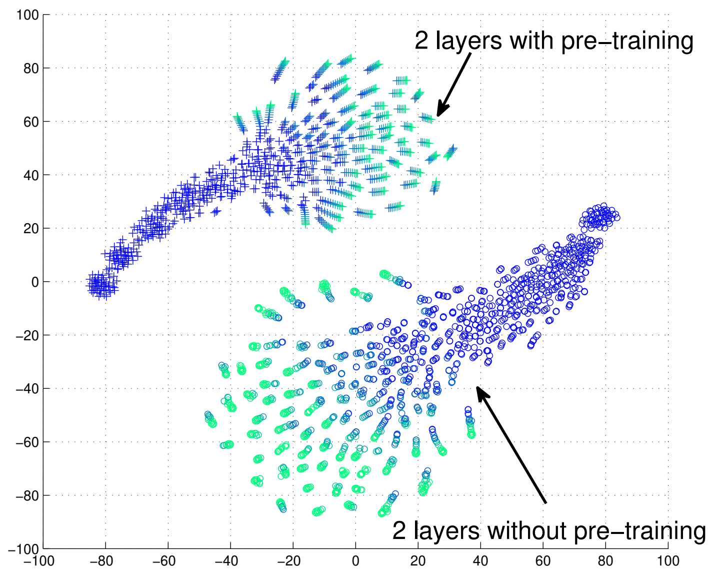

Figure 5: 2D visualizations with tSNE of the functions represented by 50 networks with and 50 networks without pre-training, as supervised training proceeds over MNIST. See Section 6.3 for an explanation. Color from dark blue to cyan and red indicates a progression in training iterations (training is longer without pre-training). The plot shows models with 2 hidden layers but results are similar with other depths.

What seems to come out of these visualizations is the following:

1. The pre-trained and not pre-trained models start and *stay* in different regions of function space.

2. From the visualization focusing on local structure (Figure 5) we see that all trajectories of a given type (with pre-training or without) initially move together. However, at some point (after about 7 epochs) the different trajectories (corresponding to different random seeds) diverge (slowing down into elongated jets) and never get back close to each other (this is more true for trajectories of networks without pre-training). This suggests that each trajectory moves into a different apparent local minimum.[11]

9. t-Distributed Stochastic Neighbor Embedding, or tSNE, by van der Maaten and Hinton (2008), with the default parameters available in the public implementation: `http://ict.ewi.tudelft.nl/~lvandermaaten/t-SNE.html`.
10. Isomap by Tenenbaum et al. (2000), with one connected component.
11. One may wonder if the divergence points correspond to a turning point in terms of overfitting. As shall be seen in Figure 8, the test error does not improve much after the 7th epoch, which reinforces this hypothesis.

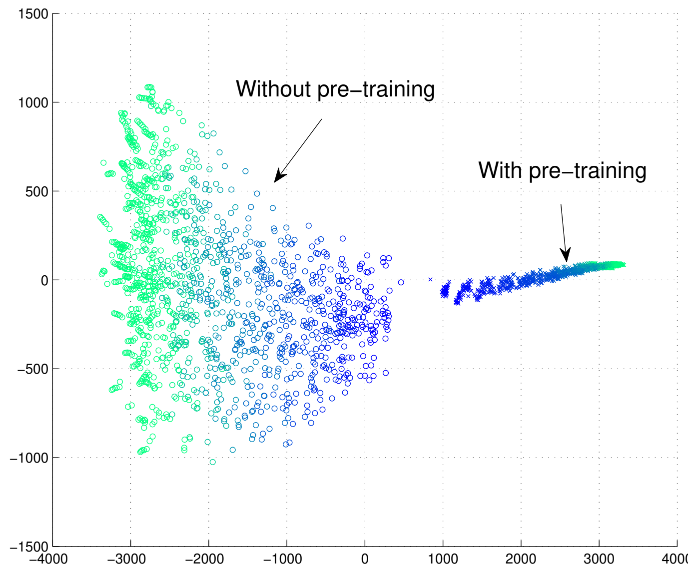

Figure 6: 2D visualization with ISOMAP of the functions represented by 50 networks with and 50 networks without pre-training, as supervised training proceeds over MNIST. See Section 6.3 for an explanation. Color from dark blue to cyan indicates a progression in training iterations (training is longer without pre-training). The plot shows models with 2 hidden layers but results are similar with other depths.

3. From the visualization focusing on global structure (Figure 6), we see the pre-trained models live in a disjoint and much smaller region of space than the not pre-trained models. In fact, from the standpoint of the functions found without pre-training, the pre-trained solutions look all the same, and their self-similarity increases (variance across seeds decreases) during training, while the opposite is observed without pre-training. This is consistent with the formalization of pre-training from Section 3, in which we described a theoretical justification for viewing unsupervised pre-training as a regularizer; there, the probabilities of pre-traininig parameters landing in a basin of attraction is small.

The visualizations of the training trajectories do seem to confirm our suspicions. It is difficult to guarantee that each trajectory actually does end up in a different local minimum (corresponding to a different function and not only to different parameters). However, all tests performed (visual inspection of trajectories in function space, but also estimation of second derivatives in the directions of all the estimated eigenvectors of the Jacobian not reported in details here) were consistent with that interpretation.

We have also analyzed models obtained at the end of training, to visualize the training criterion in the neighborhood of the parameter vector θ^* obtained. This is achieved by randomly sampling a direction v (from the stochastic gradient directions) and by plotting the training criterion around

θ^* in that direction, that is, at $\theta = \theta^* + \alpha v$, for $\alpha \in \{-2.5, -2.4, \ldots, -0.1, 0, 0.1, \ldots 2.4, 2.5\}$, and v normalized ($||v|| = 1$). This analysis is visualized in Figure 7. The error curves look close to quadratic and we seem to be near a local minimum in all directions investigated, as opposed to a saddle point or a plateau. A more definite answer could be given by computing the full Hessian eigenspectrum, which would be expensive. Figure 7 also suggests that the error landscape is a bit flatter in the case of unsupervised pre-training, and flatter for deeper architectures.

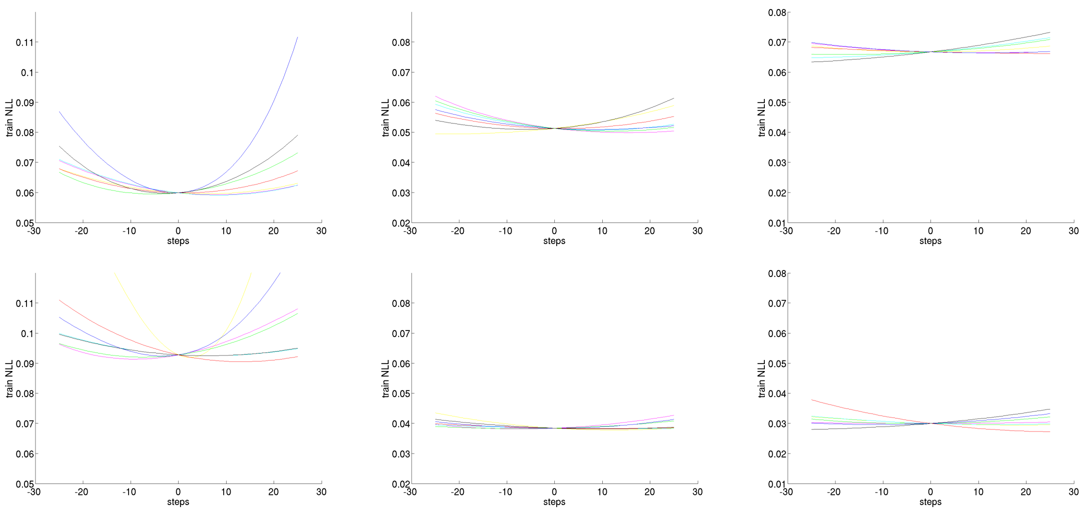

Figure 7: Training errors obtained on `Shapeset` when stepping in parameter space around a converged model in 7 random gradient directions (stepsize of 0.1). **Top**: no pre-training. **Bottom**: with unsupervised pre-training. **Left**: 1 hidden layer. **Middle**: 2 hidden layers. **Right**: 3 hidden layers. Compare also with Figure 8, where 1-layer networks with unsupervised pre-training obtain higher training errors.

6.4 Implications

The series of results presented so far show a picture that is consistent with our hypothesis. Better generalization that seems to be robust to random initializations is indeed achieved by pre-trained models, which indicates that unsupervised learning of $P(X)$ is helpful in learning $P(Y|X)$. The function space landscapes that we visualized point to the fact that there are many apparent local minima. The pre-trained models seem to end up in distinct regions of these error landscapes (and, implicitly, in different parts of the parameter space). This is both seen from the function space trajectories and from the fact that the visualizations of the learned features are qualitatively very different from those obtained by models without pre-training.

7. The Role of Unsupervised Pre-training

The observations so far in this paper confirm that starting the supervised optimization from pre-trained weights rather than from randomly initialized weights consistently yields better performing classifiers on MNIST. To better understand where this advantage came from, it is important to realize that the *supervised objective being optimized is exactly the same in both cases*. The gradient-based optimization procedure is also the same. The only thing that differs is the starting point in parameter space: either picked at random or obtained after unsupervised pre-training (which also starts from a random initialization).

Deep architectures, since they are built from the composition of several layers of non-linearities, yield an error surface that is non-convex and hard to optimize, with the suspected presence of many local minima (as also shown by the above visualizations). A gradient-based optimization should thus end in the apparent local minimum of whatever *basin of attraction* we started from. From this perspective, the advantage of unsupervised pre-training could be that it puts us in a region of parameter space where basins of attraction run deeper than when picking starting parameters at random. The advantage would be due to a better **optimization**.

Now it might also be the case that unsupervised pre-training puts us in a region of parameter space in which training error is not necessarily better than when starting at random (or possibly worse), but which systematically yields better generalization (test error). Such behavior would be indicative of a **regularization** effect. Note that the two forms of explanation are *not necessarily mutually exclusive*.

Finally, a very simple explanation could be the most obvious one: namely the disparity in the magnitude of the weights (or more generally, the marginal distribution of the weights) at the start of the supervised training phase. We shall analyze (and rule out) this hypothesis first.

7.1 Experiment 1: Does Pre-training Provide a Better Conditioning Process for Supervised Learning?

Typically gradient descent training of the deep model is initialized with randomly assigned weights, small enough to be in the linear region of the parameter space (close to zero for most neural network and DBN models). It is reasonable to ask if the advantage imparted by having an initial unsupervised pre-training phase is simply due to the weights being larger and therefore somehow providing a better "conditioning" of the initial values for the optimization process; we wanted to rule out this possibility.

By conditioning, we mean the range and marginal distribution from which we draw initial weights. In other words, could we get the same performance advantage as unsupervised pre-training if we were still drawing the initial weights independently, but from a more suitable distribution than the uniform$[-1/\sqrt{k}, 1/\sqrt{k}]$? To verify this, we performed unsupervised pre-training, and computed marginal histograms for each layer's pre-trained weights and biases (one histogram per each layer's weights and biases). We then resampled new "initial" random weights and biases according to these histograms (independently for each parameter), and performed fine-tuning from there. The resulting parameters have the same marginal statistics as those obtained after unsupervised pre-training, but not the same joint distribution.

Two scenarios can be imagined. In the first, the initialization from marginals would lead to significantly better performance than the standard initialization (when no pre-training is used). This would mean that unsupervised pre-training does provide a better marginal conditioning of the weights. In the second scenario, the marginals would lead to performance similar to or worse than that without pre-training.[12]

initialization.	Uniform	Histogram	Unsup.pre-tr.
1 layer	1.81 ± 0.07	1.94 ± 0.09	1.41 ± 0.07
2 layers	1.77 ± 0.10	1.69 ± 0.11	1.37 ± 0.09

Table 1: Effect of various initialization strategies on 1 and 2-layer architectures: independent uniform densities (one per parameter), independent densities from the marginals after unsupervised pre-training, or unsupervised pre-training (which samples the parameters in a highly dependent way so that they collaborate to make up good denoising auto-encoders.) Experiments on `MNIST`, numbers are mean and standard deviation of test errors (across different initialization seeds).

What we observe in Table 1 seems to fall within the first scenario. However, while initializing the weights to match the marginal distributions at the end of pre-training appears to slightly improve the generalization error on MNIST for 2 hidden layers, the difference is not significant and it is far from fully accounting for the discrepancy between the pre-trained and non-pre-trained results.

This experiment constitutes evidence against the preconditioning hypothesis, but does not exclude either the optimization hypothesis or the regularization hypothesis.

7.2 Experiment 2: The Effect of Pre-training on Training Error

The optimization and regularization hypotheses diverge on their prediction on how unsupervised pre-training should affect the training error: the former predicts that unsupervised pre-training should result in a lower training error, while the latter predicts the opposite. To ascertain the influence of these two possible explanatory factors, we looked at the test cost (Negative Log Likelihood on test data) obtained as a function of the training cost, along the trajectory followed in parameter space by the optimization procedure. Figure 8 shows 400 of these curves started from a point in parameter space obtained from random initialization, that is, without pre-training (blue), and 400 started from pre-trained parameters (red).

The experiments were performed for networks with 1, 2 and 3 hidden layers. As can be seen in Figure 8, while for 1 hidden layer, unsupervised pre-training reaches lower training cost than no pre-training, hinting towards a better optimization, this is not necessarily the case for the deeper networks. The remarkable observation is rather that, *at a same training cost level, the pre-trained models systematically yield a lower test cost* than the randomly initialized ones. The advantage appears to be one of *better generalization rather than merely a better optimization procedure*.

This brings us to the following result: unsupervised pre-training appears to have a similar effect to that of a good regularizer or a good "prior" on the parameters, even though no explicit regularization term is apparent in the cost being optimized. As we stated in the hypothesis, it might be reasoned that restricting the possible starting points in parameter space to those that minimize the unsupervised pre-training criterion (as with the SDAE), does in effect restrict the set of possible

12. We observed that the distribution of weights after unsupervised pre-training is fat-tailed. It is conceivable that sampling from such a distribution in order to initialize a deep architecture might actually *hurt* the performance of a deep architecture (compared to random initialization from a uniform distribution).

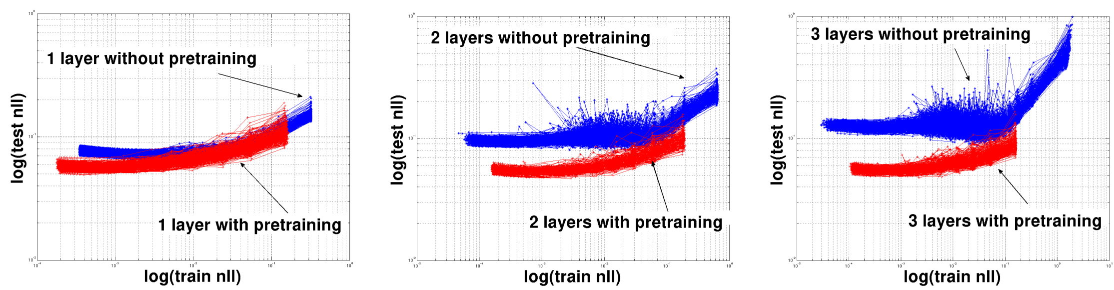

Figure 8: Evolution without pre-training (blue) and with pre-training (red) on MNIST of the log of the test NLL plotted against the log of the train NLL as training proceeds. Each of the 2×400 curves represents a different initialization. The errors are measured after each pass over the data. The rightmost points were measured after the first pass of gradient updates. Since training error tends to decrease during training, the trajectories run from right (high training error) to left (low training error). Trajectories moving up (as we go leftward) indicate a form of overfitting. All trajectories are plotted on top of each other.

final configurations for parameter values. Like regularizers in general, unsupervised pre-training (in this case, with denoising auto-encoders) might thus be seen as decreasing the variance and introducing a bias (towards parameter configurations suitable for performing denoising). Unlike ordinary regularizers, unsupervised pre-training does so in a data-dependent manner.

7.3 Experiment 3: The Influence of the Layer Size

Another signature characteristic of regularization is that the effectiveness of regularization increases as capacity (e.g., the number of hidden units) increases, effectively trading off one constraint on the model complexity for another. In this experiment we explore the relationship between the number of units per layer and the effectiveness of unsupervised pre-training. The hypothesis that unsupervised pre-training acts as a regularizer would suggest that we should see a trend of increasing effectiveness of unsupervised pre-training as the number of units per layer are increased.

We trained models on MNIST with and without pre-training using increasing layer sizes: 25, 50, 100, 200, 400, 800 units per layer. Results are shown in Figure 9. Qualitatively similar results were obtained on *Shapeset*. In the case of SDAE, we were expecting the denoising pre-training procedure to help classification performance most for large layers; this is because the denoising pre-training allows useful representations to be learned in the over-complete case, in which a layer is larger than its input (Vincent et al., 2008). What we observe is a more systematic effect: while unsupervised pre-training helps for larger layers and deeper networks, it also appears to hurt for too small networks.

Figure 9 also shows that DBNs behave qualitatively like SDAEs, in the sense that unsupervised pre-training architectures with smaller layers hurts performance. Experiments on InfiniteMNIST reveal results that are qualitatively the same. Such an experiment seemingly points to a re-verification of the regularization hypothesis. In this case, it would seem that unsupervised pre-training acts as an additional regularizer for both DBN and SDAE models—on top of the regularization provided by

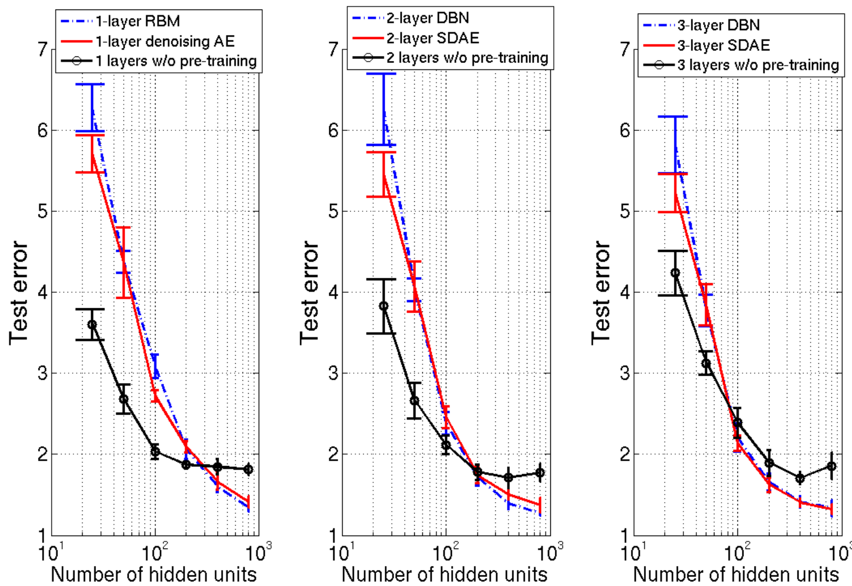

Figure 9: Effect of layer size on the changes brought by unsupervised pre-training, for networks with 1, 2 or 3 hidden layers. Experiments on MNIST. Error bars have a height of two standard deviations (over initialization seed). Pre-training hurts for smaller layer sizes and shallower networks, but it helps for all depths for larger networks.

the small size of the hidden layers. As the model size decreases from 800 hidden units, the generalization error increases, and it increases more with unsupervised pre-training presumably because of the extra regularization effect: small networks have a limited capacity already so further restricting it (or introducing an additional bias) can harm generalization. Such a result seems incompatible with a pure optimization effect. We also obtain the result that DBNs and SDAEs seem to have qualitatively similar effects as pre-training strategies.

The effect can be explained in terms of the role of unsupervised pre-training as promoting input transformations (in the hidden layers) that are useful at capturing the main variations in the input distribution $P(X)$. It may be that only a small subset of these variations are relevant for predicting the class label Y. When the hidden layers are small it is less likely that the transformations for predicting Y are included in the lot learned by unsupervised pre-training.

7.4 Experiment 4: Challenging the Optimization Hypothesis

Experiments 1–3 results are consistent with the regularization hypothesis and Experiments 2–3 would appear to directly support the regularization hypothesis over the alternative—that unsupervised pre-training aids in optimizing the deep model objective function.

In the literature there is some support for the optimization hypothesis. Bengio et al. (2007) constrained the top layer of a deep network to have 20 units and measured the training error of networks with and without pre-training. The idea was to prevent the networks from overfitting the training error simply with the top hidden layer, thus to make it clearer whether some optimization effect (of the lower layers) was going on. The reported training and test errors were lower for pre-trained networks. One problem with the experimental paradigm used by Bengio et al. (2007) is their use of early stopping. This is problematic because, as previously mentioned, early stopping is itself a regularizer, and it can influence greatly the training error that is obtained. It is conceivable that if Bengio et al. (2007) had run the models to convergence, the results could have been different. We needed to verify this.

Figure 10 shows what happens without early stopping. The training error is still higher for pre-trained networks even though the generalization error is lower. This result now favors the regularization hypothesis against the optimization story. What may have happened is that early stopping prevented the networks without pre-training from moving too much towards their apparent local minimum.

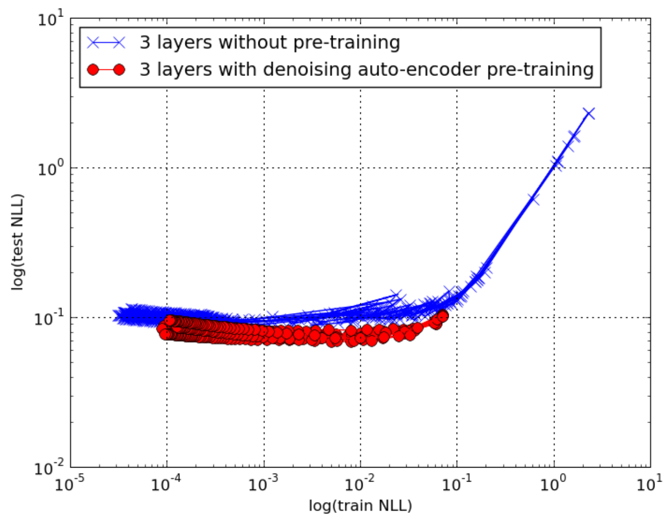

Figure 10: For MNIST, a plot of the log(train NLL) vs. log(test NLL) at each epoch of training. The top layer is constrained to 20 units.

7.5 Experiment 5: Comparing pre-training to L_1 and L_2 regularization

An alternative hypothesis would be that classical ways of regularizing could perhaps achieve the same effect as unsupervised pre-training. We investigated the effect of L_1 and L_2 regularization (i.e., adding a $||\theta||_1$ or $||\theta||_2^2$ term to the supervised objective function) in a network without pre-training. We found that while in the case of MNIST a small penalty can in principle help, the gain is nowhere near as large as it is with pre-training. For InfiniteMNIST, the optimal amount of L_1 and L_2 regularization is zero.[13]

13. Which is consistent with the classical view of regularization, in which its effect should diminish as we add more and more data.

This is not an entirely surprising finding: not all regularizers are created equal and these results are consistent with the literature on semi-supervised training that shows that unsupervised learning can be exploited as a particularly effective form of regularization.

7.6 Summary of Findings: Experiments 1-5

So far, the results obtained from the previous experiments point towards a pretty clear explanation of the effect of unsupervised pre-training: namely, that its effect is a regularization effect. We have seen that it is not simply sufficient to sample random weights with the same magnitude: the (data-dependent) unsupervised initialization is crucial. We have also observed that canonical regularizers (L_1/L_2 penalties on the weights) do not achieve the same level of performance.

The most compelling pieces of evidence in support of the regularization hypothesis are Figures 8 and 9. The alternative explanation—that unsupervised pre-training has an optimization effect—suggested by Bengio et al. (2007) doesn't seem to be supported by our experimental setup.

8. The Online Learning Setting

Our hypothesis included not only the statistical/phenomenological hypothesis that unsupervised pre-training acted as a regularizer, but also contains a mechanism for how such behavior arises both as a consequence of the dynamic nature of training—following a stochastic gradient through two phases of training and as a consequence of the non-convexity of the supervised objective function.

In our hypothesis, we posited that early examples induce changes in the magnitude of the weights that increase the amount of non-linearity of the network, which in turn decreases the number of regions accessible to the stochastic gradient descent procedure. This means that the early examples (be they pre-training examples or otherwise) determine the basin of attraction for the remainder of training; this also means that the early examples have a disproportionate influence on the configuration of parameters of the trained models.

One consequence to the hypothesized mechanism is that we would predict that in the online learning setting with unbounded or very large data sets, the behavior of unsupervised pre-training would diverge from the behavior of a canonical regularizer (L_1/L_2). This is because the effectiveness of a canonical regularizer **decreases** as the data set grows, whereas the effectiveness of unsupervised pre-training as a regularizer is **maintained** as the data set grows.

Note that stochastic gradient descent in online learning is a stochastic gradient descent optimization of the generalization error, so good online error in principle implies that we are optimizing well the generalization error. Indeed, each gradient $\frac{\partial L(x,y)}{\partial \theta}$ for example (x,y) (with $L(x,y)$ the supervised loss with input x and label y) sampled from the true generating distribution $P(x,y)$ is an unbiased Monte-Carlo estimator of the true gradient of generalization error, that is, $\sum_y \int_x \frac{\partial L(x,y)}{\partial \theta} P(x,y) dx$.

In this section we empirically challenge this aspect of the hypothesis and show that the evidence does indeed support our hypothesis over what is more typically expected from a regularizer.

8.1 Experiment 6: Effect of Pre-training with Very Large Data Sets

The results presented here are perhaps the most surprising findings of this paper. Figure 11 shows the online classification error (on the next block of examples, as a moving average) for 6 architectures that are trained on `InfiniteMNIST`: 1 and 3-layer DBNs, 1 and 3-layer SDAE, as well as 1 and 3-layer networks without pre-training.

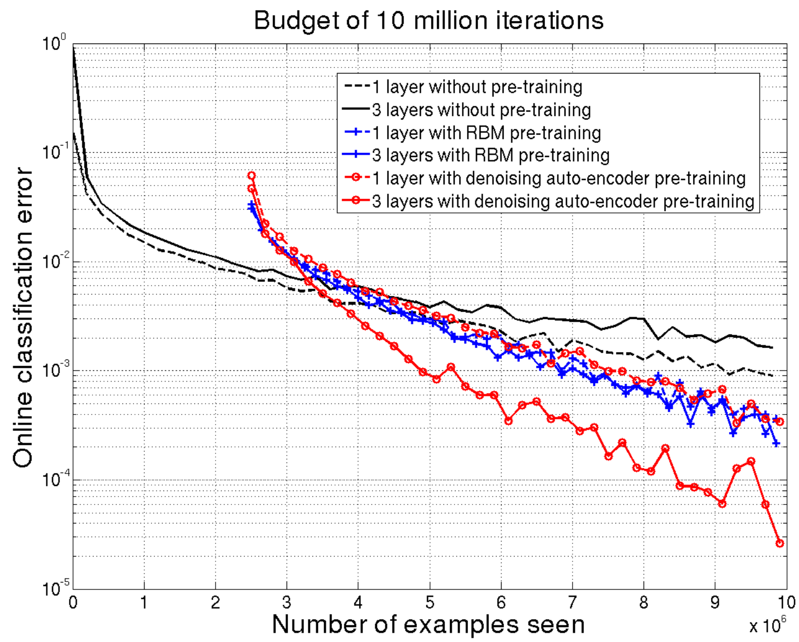

Figure 11: Comparison between 1 and 3-layer networks trained on `InfiniteMNIST`. Online classification error, computed as an average over a block of last 100,000 errors.

We can draw several observations from these experiments. First, 3-layer networks without pre-training are worse at generalization, compared to the 1-layer equivalent. This confirms the hypothesis that even in an online setting, optimization of deep networks is harder than shallow ones. Second, 3-layer SDAE models seem to generalize better than 3-layer DBNs. Finally and most importantly, the pre-training advantage does not vanish as the number of training examples increases, on the contrary.

Note that the number of hidden units of each model is a hyperparameter.[14] So theoretical results suggest that 1-layer networks without pre-training should in principle be able to represent the input distribution as capacity and data grow. Instead, without pre-training, the networks are not able to take advantage of the additional capacity, which again points towards the optimization explanation. It is clear, however, that **the starting point of the non-convex optimization matters**, even for networks that are seemingly "easier" to optimize (1-layer ones), which supports our hypothesis.

Another experiment that shows the effects of large-scale online stochastic non-convex optimization is shown in Figure 12. In the setting of `InfiniteMNIST`, we compute the error on the *training set*, in the same order that we presented the examples to the models. We observe several interesting results: first, note that both models are better at classifying more recently seen examples. This is a natural effect of stochastic gradient descent with a constant learning rate (which gives exponentially more weight to recent examples). Note also that examples at the beginning of training are essentially like test examples for both models, in terms of error. Finally, we observe that the pre-trained

14. This number was chosen individually for each model s.t. the error on the last 1 million examples is minimized. In practice, this meant 2000 units for 1-layer networks and 1000 units/layer for 3-layer networks.

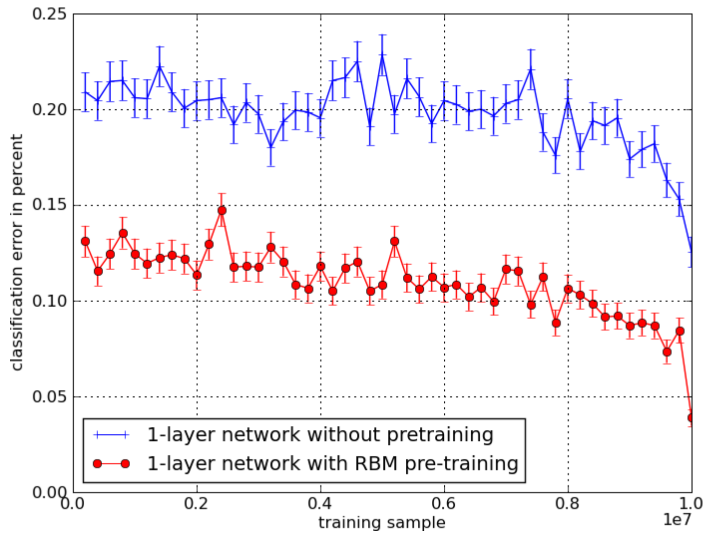

Figure 12: Error of 1-layer network with RBM pre-training and without, on the 10 million examples used for training it. The errors are calculated in the same order (from left to right, above) as the examples were presented during training. Each error bar corresponds to a block of consecutive training examples.

model is better across the board *on the training set*. This fits well with the optimization hypothesis, since it shows that unsupervised pre-training has an optimization effect.

What happens in this setting is that the training and generalization errors converge as the empirical distribution (defined by the training set) converges to the true data distribution. These results show that the effectiveness of unsupervised pre-training does not diminish with increasing data set sizes. This would be unexpected from a superficial understanding of unsupervised pre-training as a regularization method. However it is entirely consistent with our interpretation, stated in our hypothesis, of the role of unsupervised pre-training in the online setting with stochastic gradient descent training on a non-convex objective function.

8.2 Experiment 7: The Effect of Example Ordering

The hypothesized mechanism implies, due to the dynamics of learning—the increase in weight magnitude and non-linearity as training proceeds, as well as the dependence of the basin of attraction on early data—that, when training with stochastic gradient descent, we should see increased sensitivity to early examples. In the case of `InfiniteMNIST` we operate in an online stochastic optimization regime, where we try to find a local minimum of a highly non-convex objective function. It is then interesting to study to what extent the outcome of this optimization is influenced by the examples seen at different points during training, and whether the early examples have a stronger influence (which would not be the case with a convex objective).

To quantify the variance of the outcome with respect to training samples at different points during training, and to compare these variances for models with and without pre-training, we proceeded with the following experiment. Given a data set with 10 million examples, we vary (by resampling) the first million examples (across 10 different random draws, sampling a different set of 1 million examples each time) and keep the other ones fixed. After training the (10) models, we measure the variance (across the 10 draws) of the *output* of the networks on a fixed test set (i.e., we measure the variance in function space). We then vary the next million examples in the same fashion, and so on, to see how much each of the ten parts of the training set influenced the final function.

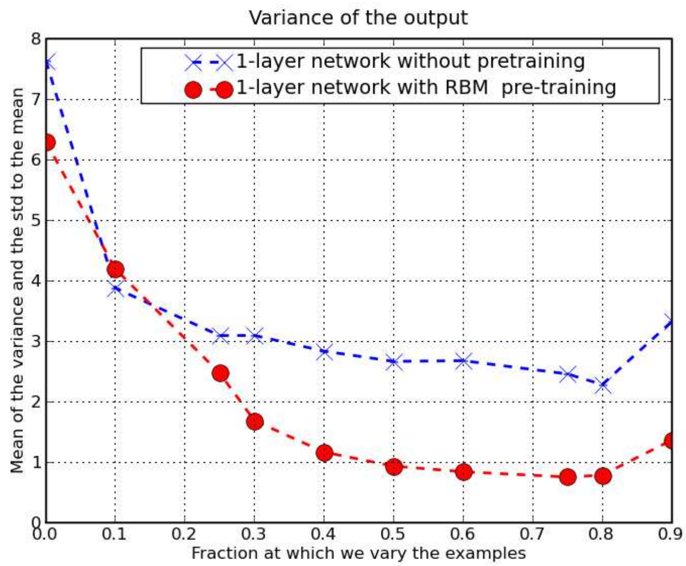

Figure 13: Variance of the output of a trained network with 1 layer. The variance is computed as a function of the point at which we vary the training samples. Note that the 0.25 mark corresponds to the start of pre-training.

Figure 13 shows the outcome of such an analysis. The samples at the beginning[15] do seem to influence the output of the networks more than the ones at the end. However, this variance is *lower* for the networks that have been pre-trained. In addition to that, one should note that the variance of pre-trained network at 0.25 (i.e., the variance of the output as a function of the first samples used for supervised training) is *lower* than the variance of the supervised network at 0.0. Such results imply that unsupervised pre-training can be seen as a sort of variance reduction technique, consistent with a regularization hypothesis. Finally, both networks are more influenced by the *last examples* used for optimization, which is simply due to the fact that we use stochastic gradient with a constant learning rate, where the most recent examples' gradient has a greater influence.

These results are consistent with what our hypothesis predicts: both the fact that early examples have greater influence (i.e., the variance is higher) and that pre-trained models seem to reduce this variance are in agreement with what we would have expected.

15. Which are *unsupervised* examples, for the red curve, until the 0.25 mark in Figure 13.

8.3 Experiment 8: Pre-training only k layers

From Figure 11 we can see that unsupervised pre-training makes quite a difference for 3 layers, on `InfiniteMNIST`. In Figure 14 we explore the link between depth and unsupervised pre-training in more detail. The setup is as follows: for both `MNIST` and `InfiniteMNIST` we pre-train only the bottom k layers and randomly initialize the top $n-k$ layers in the usual way. In this experiment, $n=3$ and we vary k from 0 (which corresponds to a network with no pre-training) to $k=n$ (which corresponds to the normal pre-trained case).

For `MNIST`, we plot the log(train NLL) vs. log(test NLL) trajectories, where each point corresponds to a measurement after a certain number of epochs. The trajectories go roughly from the right to left and from top to bottom, corresponding to the lowering of the training and test errors. We can also see that models overfit from a certain point onwards.

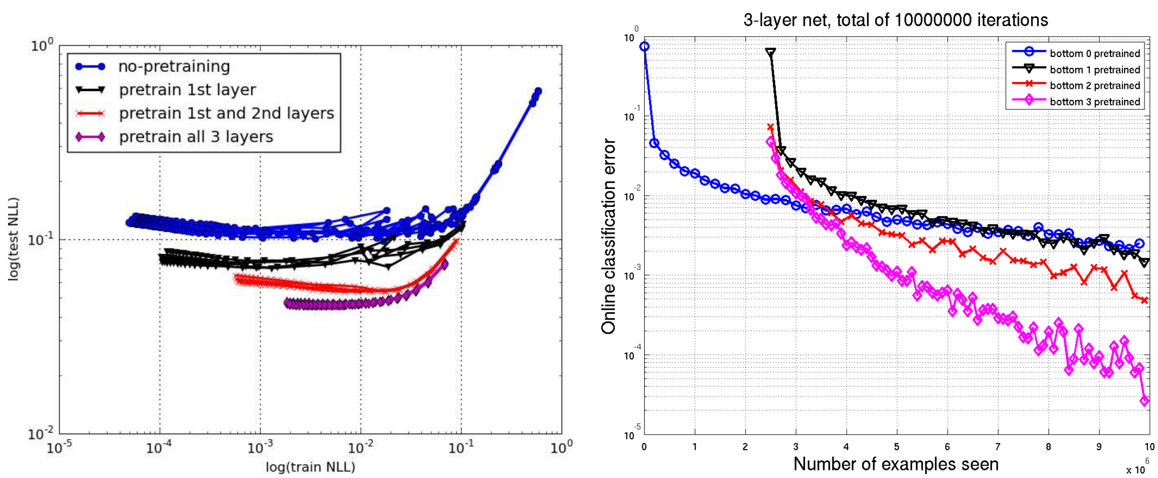

Figure 14: *On the left*: for `MNIST`, a plot of the log(train NLL) vs. log(test NLL) at each epoch of training. We pre-train the first layer, the first two layers and all three layers using RBMs and randomly initialize the other layers; we also compare with the network whose layers are all randomly initialized. *On the right*: `InfiniteMNIST`, the online classification error. We pre-train the first layer, the first two layers or all three layers using denoising auto-encoders and leave the rest of the network randomly initialized.

For `InfiniteMNIST`, we simply show the online error. The results are ambiguous w.r.t the difficulty of optimizing the lower layers versus the higher ones. We would have expected that the largest incremental benefit came from pre-training the first layer or first two layers. It is true for the first two layers, but not the first. As we pre-train more layers, the models become better at generalization. In the case of the finite `MNIST`, note how the final training error (after the same number of epochs) becomes *worse* with pre-training of more layers. This clearly brings additional support to the regularization explanation.

9. Discussion and Conclusions

We have shown that unsupervised pre-training adds robustness to a deep architecture. The same set of results also suggests that increasing the depth of an architecture that is not pre-trained increases the probability of finding poor apparent local minima. Pre-trained networks give consistently better generalization. Our visualizations point to the observations that pre-trained networks learn qualitatively different features (if networks are visualized in the weight space) compared to networks without pre-training. Moreover, the trajectories of networks with different initialization seeds seem to fall into many distinct apparent local minima, which are again different (and seemingly far apart) depending on whether we use pre-training or not.

We have shown that unsupervised pre-training is not simply a way of getting a good initial marginal distribution, and that it captures more intricate dependencies between parameters. One of our findings is that deep networks with unsupervised pre-training seem to exhibit some properties of a regularizer: with small enough layers, pre-trained deep architectures are systematically worse than randomly initialized deep architectures. Moreover, when the layers are big enough, the pre-trained models obtain worse training errors, but better generalization performance. Additionally, we have re-done an experiment which purportedly showed that unsupervised pre-training can be explained with an optimization hypothesis and observed a regularization effect instead. We also showed that classical regularization techniques (such as L_1/L_2 penalties on the network weights) cannot achieve the same performance as unsupervised pre-training, and that the effect of unsupervised pre-training does not go away with more training data, so if unsupervised pre-training is a regularizer, it is certainly of a rather different kind.

The two unsupervised pre-training strategies considered—denoising auto-encoders and Restricted Boltzmann Machines—seem to produce qualitatively similar observations. We have observed that, surprisingly, the pre-training advantage is present even in the case of really large training sets, pointing towards the conclusion that the starting point in the non-convex optimization problem is indeed quite important; a fact confirmed by our visualizations of filters at various levels in the network. Finally, the other important set of results show that unsupervised pre-training acts like a variance reduction technique, yet a network with pre-training has a lower training error on a very large data set, which supports an optimization interpretation of the effect of pre-training.

How do we make sense of all these results? The contradiction between what looks like regularization effects and what looks like optimization effects appears, on the surface, unresolved. Instead of sticking to these labels, we attempted to draw a hypothesis, described in Section 3 about the dynamics of learning in an architecture that is trained using two phases (unsupervised pre-training and supervised fine-tuning), which we believe to be consistent with all the above results.

This hypothesis suggests that there are consequences of the non-convexity of the supervised objective function, which we observed in various ways throughout our experiments. One of these consequences is that early examples have a big influence on the outcome of training and this is one of the reasons why in a large-scale setting the influence of unsupervised pre-training is still present. Throughout this paper, we have delved on the idea that the basin of attraction induced by the early examples (in conjunction with unsupervised pre-training) is, for all practical purposes, a basin from which supervised training does not escape.

This effect can be observed from the various visualizations and performance evaluations that we made. *Unsupervised pre-training,* **as a regularizer** *that only influences the starting point of supervised training, has an effect that, contrary to classical regularizers, does not disappear with* *more data* (at least as far as we can see from our results). Basically, unsupervised pre-training favors hidden units that compute features of the input X that correspond to major factors of variation in the true $P(X)$. Assuming that some of these are near features useful at predicting variations in Y, unsupervised pre-training sets up the parameters near a solution of low predictive generalization error.

One of the main messages that our results imply is that the optimization of a non-convex objective function with stochastic gradient descent presents challenges for analysis, especially in a regime with large amounts of data. Our analysis so far shows that it is possible for networks that are trained in such a regime to be influenced more by early examples. This can pose problems in scenarios where we would like our networks to be able to capture more of the information in later examples, that is, when training from very large data sets and trying to capture a lot of information from them.

One interesting realization is that with a small training set, we do not usually put a lot of importance on minimizing the training error, because overfitting is a major issue; the training error is not a good way to distinguish between the generalization performance of two models. In that setting, unsupervised pre-training helps to find apparent local minima that have better generalization error. With a large training set, as we saw in Figure 12, the empirical and true distributions converge. In such a scenario, *finding a better apparent local minimum will matter and stronger (better) optimization strategies should have a significant impact on generalization when the training set is very large*. Note also that it would be interesting to extend our experimental techniques to the problem of training deep auto-encoders (with a bottleneck), where previous results (Hinton and Salakhutdinov, 2006) show that not only test error but also training error is greatly reduced by unsupervised pre-training, which is a strong indicator of an optimization effect. We hypothesize that the presence of the bottleneck is a crucial element that distinguishes the deep auto-encoders from the deep classifiers studied here.

In spite of months of CPU time on a cluster devoted to the experiments described here (which is orders of magnitude more than most previous work in this area), more could certainly be done to better understand these effects. Our original goal was to have well-controlled experiments with well understood data sets. It was not to advance a particular algorithm but rather to try to better understand a phenomenon that has been well documented elsewhere. Nonetheless, our results are limited by the data sets used and it is plausible that different conclusions could be drawn, should the same experiments be carried out on other data.

Our results suggest that optimization in deep networks is a complicated problem that is influenced in great part by the early examples during training. Future work should clarify this hypothesis. If it is true and we want our learners to capture really complicated distributions from very large training sets, it may mean that we should consider learning algorithms that reduce the effect of the early examples, allowing parameters to escape from the attractors in which current learning dynamics get stuck.

The observations reported here suggest more detailed explanations than those already discussed, which could be tested in future work. We hypothesize that the factors of variation present in the input distribution are disentangled more and more as we go from the input layer to higher-levels of the feature hierarchy. This is coherent with observations of increasing invariance to geometric transformations in DBNs trained on images (Goodfellow et al., 2009), as well as by visualizing the variations in input images generated by sampling from the model (Hinton, 2007; Susskind et al., 2008), or when considering the preferred input associated with different units at different depths (Lee et al., 2009; Erhan et al., 2009). As a result, during early stages of learning, the upper layers (those that typically learn quickly) would have access to a more robust representation of the input and are less likely to be hindered by the entangling of factors variations present in the input. If this disentangling hypothesis is correct, it would help to explain how unsupervised pre-training can address the chicken-and-egg issue explained in Section 2: the lower layers of a supervised deep architecture need the upper layers to define what they should extract, and vice-versa. Instead, the lower layers can extract robust and disentangled representations of the factors of variation and the upper layers select and combine the appropriate factors (sometimes not all at the top hidden layer). Note that as factors of variation are disentangled, it could also happen that some of them are not propagated upward (before fine-tuning), because RBMs do not try to represent in their hidden layer input bits that are independent.

To further explain why smaller hidden layers yield worse performance with pre-training than without (Figure 9), one may hypothesize further that, for some data sets, the leading factors of variation present in $P(X)$ (presumably the only ones captured in a smaller layer) are less predictive of Y than random projections[16] can be, precisely because of the hypothesized disentangling effect. With enough hidden units, unsupervised pre-training may extract among the larger set of learned features some that are highly predictive of Y (more so than random projections). This additional hypothesis could be tested by measuring the mutual information between each hidden unit and the object categories (as done by Lee et al., 2009), as the number of hidden units is varied (like in Figure 9). It is expected that the unit with the most mutual information will be less informative with pre-training when the number of hidden units is too small, and more informative with pre-training when the number of hidden units is large enough.

Under the hypothesis that we have proposed in Section 3, the following result is unaccounted for: in Figure 8(a), training error is lower with pre-training when there is only one hidden layer, but worse with more layers. This may be explained by the following additional hypothesis. Although each layer extracts information about Y in some of its features, it is not guaranteed that all of that information is preserved when moving to higher layers. One may suspect this in particular for RBMs, which would not encode in their hidden layer any input bits that would be marginally independent of the others, because these bits would be explained by the visible biases: perfect disentangling of Y from the other factors of variation in X may yield marginally independent bits about Y. Although supervised fine-tuning should help to bubble up that information towards the output layer, it might be more difficult to do so for deeper networks, explaining the above-stated feature of Figure 8. Instead, in the case of a single hidden layer, less information about Y would have been dropped (if at all), making the job of the supervised output layer easier. This is consistent with earlier results (Larochelle et al., 2009) showing that for several data sets supervised fine-tuning significantly improves classification error, when the output layer only takes input from the top hidden layer. This hypothesis is also consistent with the observation made here (Figure 1) that unsupervised pre-training actually does not help (and can hurt) for too deep networks.

In addition to exploring the above hypotheses, future work should include an investigation of the connection between the results presented in this paper and by Hinton and Salakhutdinov (2006), where it seems to be hard to obtain a good training reconstruction error with deep auto-encoders (in an unsupervised setting) without performing pre-training. Other avenues for future work include the analysis and understanding of deep semi-supervised techniques where one does not separate

16. Meaning the random initialization of hidden layers.

between the pre-training phase and the supervised phase, such as work by Weston et al. (2008) and Larochelle and Bengio (2008). Such algorithms fall more squarely into the realm of semi-supervised methods. We expect that analyses similar to the ones we performed would be potentially harder, but perhaps revealing as well.

Many open questions remain towards understanding and improving deep architectures. Our conviction is that devising improved strategies for learning in deep architectures requires a more profound understanding of the difficulties that we face with them. This work helps with such understanding via extensive simulations and puts forward a hypothesis explaining the mechanisms behind unsupervised pre-training, which is well supported by our results.

Acknowledgments

This research was supported by funding from NSERC, MITACS, FQRNT, and the Canada Research Chairs. The authors also would like to thank the editor and reviewers, as well as Fernando Pereira for their helpful comments and suggestions.

References

Shun-ichi Amari, Noboru Murata, Klaus-Robert Müller, Michael Finke, and Howard Hua Yang. Asymptotic statistical theory of overtraining and cross-validation. *IEEE Transactions on Neural Networks*, 8(5):985–996, 1997.

Lalit Bahl, Peter Brown, Peter deSouza, and Robert Mercer. Maximum mutual information estimation of hidden markov parameters for speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 49–52, Tokyo, Japan, 1986.

Andrew E. Barron. Complexity regularization with application to artificial neural networks. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 561–576. Kluwer Academic Publishers, 1991.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS'01)*, Cambridge, MA, 2002. MIT Press.

Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009. Also published as a book. Now Publishers, 2009.

Yoshua Bengio and Olivier Delalleau. Justifying and generalizing contrastive divergence. *Neural Computation*, 21(6):1601–1621, June 2009.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Machines*, pages 321–360. MIT Press, 2007.

Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. The curse of highly variable functions for local kernel machines. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18 (NIPS'05)*, pages 107–114. MIT Press, Cambridge, MA, 2006.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In Bernhard Schölkopf, John Platt, and Thomas Hoffman, editors, *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pages 153–160. MIT Press, 2007.

Marc H. Bornstein. *Sensitive periods in development : interdisciplinary perspectives / edited by Marc H. Bornstein*. Lawrence Erlbaum Associates, Hillsdale, N.J. :, 1987.

Olivier Chapelle, Jason Weston, and Bernhard Schölkopf. Cluster kernels for semi-supervised learning. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15 (NIPS'02)*, pages 585–592, Cambridge, MA, 2003. MIT Press.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, pages 160–167. ACM, 2008.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, Université de Montréal, 2009.

Patrick Gallinari, Yann LeCun, Sylvie Thiria, and Francoise Fogelman-Soulie. Memoires associatives distribuees. In *Proceedings of COGNITIVA 87*, Paris, La Villette, 1987.

Ian Goodfellow, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 646–654. 2009.

Raia Hadsell, Ayse Erkan, Pierre Sermanet, Marco Scoffier, Urs Muller, and Yann LeCun. Deep belief net learning in a long-range vision system for autonomous off-road driving. In *Proc. Intelligent Robots and Systems (IROS'08)*, pages 628–633, 2008.

Johan Håstad. Almost optimal lower bounds for small depth circuits. In *Proceedings of the 18th annual ACM Symposium on Theory of Computing*, pages 6–20, Berkeley, California, 1986. ACM Press.

Johan Håstad and Mikael Goldmann. On the power of small-depth threshold circuits. *Computational Complexity*, 1:113–129, 1991.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

Geoffrey E. Hinton. To recognize shapes, first learn to generate images. In Paul Cisek, Trevor Drew, and John Kalaska, editors, *Computational Neuroscience: Theoretical Insights into Brain Function*. Elsevier, 2007.

Geoffrey E. Hinton and Ruslan Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.

Goeffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

Hugo Larochelle and Yoshua Bengio. Classification using discriminative restricted Boltzmann machines. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, pages 536–543. ACM, 2008.

Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Int. Conf. Mach. Learn.*, pages 473–480, 2007.

Hugo Larochelle, Yoshua Bengio, Jerome Louradour, and Pascal Lamblin. Exploring strategies for training deep neural networks. *The Journal of Machine Learning Research*, 10:1–40, January 2009.

Julia A. Lasserre, Christopher M. Bishop, and Thomas P. Minka. Principled hybrids of generative and discriminative models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'06)*, pages 87–94, Washington, DC, USA, 2006. IEEE Computer Society.

Yann LeCun. *Modèles connexionistes de l'apprentissage*. PhD thesis, Université de Paris VI, 1987.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Honglak Lee, Chaitanya Ekanadham, and Andrew Ng. Sparse deep belief net model for visual area V2. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS'07)*, pages 873–880. MIT Press, Cambridge, MA, 2008.

Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Léon Bottou and Michael Littman, editors, *Proceedings of the Twenty-sixth International Conference on Machine Learning (ICML'09)*. ACM, Montreal (Qc), Canada, 2009.

Gaëlle Loosli, Stéphane Canu, and Léon Bottou. Training invariant support vector machines using selective sampling. In Léon Bottou, Olivier Chapelle, Dennis DeCoste, and Jason Weston, editors, *Large Scale Kernel Machines*, pages 301–320. MIT Press, Cambridge, MA., 2007.

Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep learning from temporal coherence in video. In Léon Bottou and Michael Littman, editors, *Proceedings of the 26th International Conference on Machine Learning*, pages 737–744, Montreal, June 2009. Omnipress.

Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS'01)*, pages 841–848, 2002.

Simon Osindero and Geoffrey E. Hinton. Modeling image patches with a directed hierarchy of markov random field. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS'07)*, pages 1121–1128, Cambridge, MA, 2008. MIT Press.

Dan Povey and Philip C. Woodland. Minimum phone error and i-smoothing for improved discriminative training. In *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, volume 1, pages I–105–I–108 vol.1, 2002.

Marc'Aurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann LeCun. Efficient learning of sparse representations with an energy-based model. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pages 1137–1144. MIT Press, 2007.

Marc'Aurelio Ranzato, Y-Lan Boureau, and Yann LeCun. Sparse feature learning for deep belief networks. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS'07)*, pages 1185–1192, Cambridge, MA, 2008. MIT Press.

Ruslan Salakhutdinov and Geoffrey E. Hinton. Using deep belief nets to learn covariance kernels for Gaussian processes. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS'07)*, pages 1249–1256, Cambridge, MA, 2008. MIT Press.

Ruslan Salakhutdinov and Geoffrey E. Hinton. Semantic hashing. In *Proceedings of the 2007 Workshop on Information Retrieval and applications of Graphical Models (SIGIR 2007)*, Amsterdam, 2007. Elsevier.

Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey E. Hinton. Restricted Boltzmann machines for collaborative filtering. In Zoubin Ghahramani, editor, *Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML'07)*, pages 791–798, New York, NY, USA, 2007. ACM.

Sebastian H. Seung. Learning continuous attractors in recurrent networks. In M.I. Jordan, M.J. Kearns, and S.A. Solla, editors, *Advances in Neural Information Processing Systems 10 (NIPS'97)*, pages 654–660. MIT Press, 1998.

Jonas Sjöberg and Lennart Ljung. Overtraining, regularization and searching for a minimum, with application to neural networks. *International Journal of Control*, 62(6):1391–1407, 1995.

Joshua M. Susskind, Geoffrey E., Javier R. Movellan, and Adam K. Anderson. Generating facial expressions with deep belief nets. In V. Kordic, editor, *Affective Computing, Emotion Modelling, Synthesis and Recognition*, pages 421–440. ARS Publishers, 2008.

Joshua Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, November 2008.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In Andrew McCallum and Sam Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 1096–1103. Omnipress, 2008.

Max Welling, Michal Rosen-Zvi, and Geoffrey E. Hinton. Exponential family harmoniums with an application to information retrieval. In L.K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17 (NIPS'04)*, pages 1481–1488, Cambridge, MA, 2005. MIT Press.

Jason Weston, Frédéric Ratle, and Ronan Collobert. Deep learning via semi-supervised embedding. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, pages 1168–1175, New York, NY, USA, 2008. ACM.

Andrew Yao. Separating the polynomial-time hierarchy by oracles. In *Proceedings of the 26th Annual IEEE Symposium on Foundations of Computer Science*, pages 1–10, 1985.

Long Zhu, Yuanhao Chen, and Alan Yuille. Unsupervised learning of probabilistic grammar-markov models for object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):114–128, 2009.