Learning Deep Architectures for AI

Learning Deep Architectures for AI

Yoshua Bengio

Dept. IRO, Université de Montréal
C.P. 6128, Montreal, Qc
Canada

yoshua.bengio@umontreal.ca

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in
Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is Y. Bengio, Learning Deep Architectures for AI, Foundation and Trends® in Machine Learning, vol 2, no 1, pp 1–127, 2009

ISBN: 978-1-60198-294-0

Foundations and Trends® in
Machine Learning
Volume 2 Issue 1, 2009

Editorial Scope

Foundations and Trends® in Machine Learning will publish survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2009, Volume 2, 4 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Foundations and Trends® in
Machine Learning
Vol. 2, No. 1 (2009) 1–127
© 2009 Y. Bengio
DOI: 10.1561/2200000006

Learning Deep Architectures for AI

Yoshua Bengio

Dept. IRO, Université de Montréal, C.P. 6128, Montreal, Qc, H3C 3J7, Canada, yoshua.bengio@umontreal.ca

Abstract

Theoretical results 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*. Deep architectures are composed of multiple levels of non-linear operations, such as in neural nets with many hidden layers or in complicated propositional formulae re-using many sub-formulae. Searching the parameter space of deep architectures is a difficult task, but learning algorithms such as those for Deep Belief Networks have recently been proposed to tackle this problem with notable success, beating the state-of-the-art in certain areas. This monograph discusses the motivations and principles regarding learning algorithms for deep architectures, in particular those exploiting as building blocks unsupervised learning of single-layer models such as Restricted Boltzmann Machines, used to construct deeper models such as Deep Belief Networks.

Contents

1 Introduction 1

1.1 How do We Train Deep Architectures? 4
1.2 Intermediate Representations: Sharing Features and Abstractions Across Tasks 6
1.3 Desiderata for Learning AI 9
1.4 Outline of the Paper 10

2 Theoretical Advantages of Deep Architectures 13

2.1 Computational Complexity 16
2.2 Informal Arguments 18

3 Local vs Non-Local Generalization 21

3.1 The Limits of Matching Local Templates 21
3.2 Learning Distributed Representations 27

4 Neural Networks for Deep Architectures 31

4.1 Multi-Layer Neural Networks 31
4.2 The Challenge of Training Deep Neural Networks 32
4.3 Unsupervised Learning for Deep Architectures 40
4.4 Deep Generative Architectures 41
4.5 Convolutional Neural Networks 44
4.6 Auto-Encoders 46

5 Energy-Based Models and Boltzmann Machines 49

5.1 Energy-Based Models and Products of Experts 49
5.2 Boltzmann Machines 54
5.3 Restricted Boltzmann Machines 56
5.4 Contrastive Divergence 60

6 Greedy Layer-Wise Training of Deep Architectures 69

6.1 Layer-Wise Training of Deep Belief Networks 69
6.2 Training Stacked Auto-Encoders 72
6.3 Semi-Supervised and Partially Supervised Training 73

7 Variants of RBMs and Auto-Encoders 75

7.1 Sparse Representations in Auto-Encoders and RBMs 75
7.2 Denoising Auto-Encoders 81
7.3 Lateral Connections 83
7.4 Conditional RBMs and Temporal RBMs 84
7.5 Factored RBMs 86
7.6 Generalizing RBMs and Contrastive Divergence 87

8 Stochastic Variational Bounds for Joint Optimization of DBN Layers 91

8.1 Unfolding RBMs into Infinite Directed Belief Networks 92
8.2 Variational Justification of Greedy Layer-wise Training 94
8.3 Joint Unsupervised Training of All the Layers 97

9 Looking Forward 101

9.1 Global Optimization Strategies 101
9.2 Why Unsupervised Learning is Important 107
9.3 Open Questions 108

10 Conclusion **113**

Acknowledgments **115**

References **117**

1

Introduction

Allowing computers to model our world well enough to exhibit what we call intelligence has been the focus of more than half a century of research. To achieve this, it is clear that a large quantity of information about our world should somehow be stored, explicitly or implicitly, in the computer. Because it seems daunting to formalize manually all that information in a form that computers can use to answer questions and generalize to new contexts, many researchers have turned to *learning algorithms* to capture a large fraction of that information. Much progress has been made to understand and improve learning algorithms, but the challenge of artificial intelligence (AI) remains. Do we have algorithms that can understand scenes and describe them in natural language? Not really, except in very limited settings. Do we have algorithms that can infer enough semantic concepts to be able to interact with most humans using these concepts? No. If we consider image understanding, one of the best specified of the AI tasks, we realize that we do not yet have learning algorithms that can discover the many visual and semantic concepts that would seem to be necessary to interpret most images on the web. The situation is similar for other AI tasks.

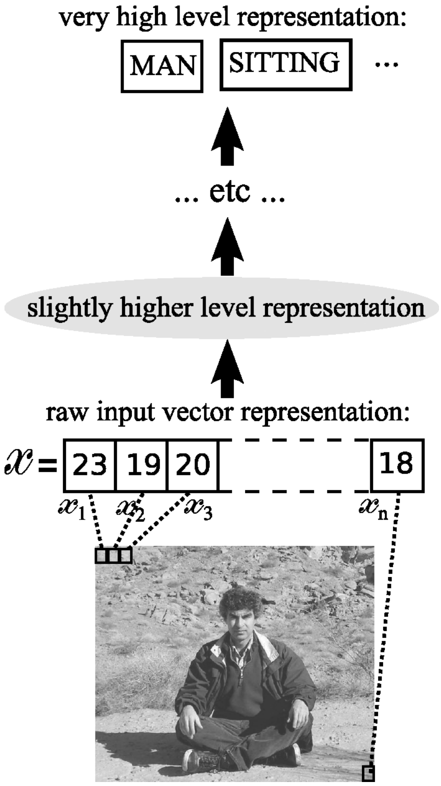
very high level representation:
MAN
SITTING
...
... etc ...
slightly higher level representation
raw input vector representation:
x =
23
19
20
18
x_1
x_2
x_3
x_n

Fig. 1.1 We would like the raw input image to be transformed into gradually higher levels of representation, representing more and more abstract functions of the raw input, e.g., edges, local shapes, object parts, etc. In practice, we do not know in advance what the "right" representation should be for all these levels of abstractions, although linguistic concepts might help guessing what the higher levels should implicitly represent.

Consider for example the task of interpreting an input image such as the one in Figure 1.1. When humans try to solve a particular AI task (such as machine vision or natural language processing), they often exploit their intuition about how to decompose the problem into subproblems and multiple levels of representation, e.g., in object parts and constellation models [138, 179, 197] where models for parts can be re-used in different object instances. For example, the current state-of-the-art in machine vision involves a sequence of modules starting from pixels and ending in a linear or kernel classifier [134, 145], with intermediate modules mixing engineered transformations and learning, e.g., first extracting low-level features that are invariant to small geometric variations (such as edge detectors from Gabor filters), transforming them gradually (e.g., to make them invariant to contrast changes and contrast inversion, sometimes by pooling and sub-sampling), and then detecting the most frequent patterns. A plausible and common way to extract useful information from a natural image involves transforming the raw pixel representation into gradually more abstract representations, e.g., starting from the presence of edges, the detection of more complex but local shapes, up to the identification of abstract categories associated with sub-objects and objects which are parts of the image, and putting all these together to capture enough understanding of the scene to answer questions about it.

Here, we assume that the computational machinery necessary to express complex behaviors (which one might label "intelligent") requires *highly varying* mathematical functions, i.e., mathematical functions that are highly non-linear in terms of raw sensory inputs, and display a very large number of variations (ups and downs) across the domain of interest. We view the raw input to the learning system as a high dimensional entity, made of many observed variables, which are related by unknown intricate statistical relationships. For example, using knowledge of the 3D geometry of solid objects and lighting, we can relate small variations in underlying physical and geometric factors (such as position, orientation, lighting of an object) with changes in pixel intensities for all the pixels in an image. We call these *factors of variation* because they are different aspects of the data that can vary separately and often independently. In this case, explicit knowledge of the physical factors involved allows one to get a picture of the mathematical form of these dependencies, and of the shape of the set of images (as points in a high-dimensional space of pixel intensities) associated with the same 3D object. If a machine captured the factors that explain the statistical variations in the data, and how they interact to generate the kind of data we observe, we would be able to say that the machine *understands* those aspects of the world covered by these factors of variation. Unfortunately, in general and for most factors of variation underlying natural images, we do not have an analytical understanding of these factors of variation. We do not have enough formalized prior knowledge about the world to explain the observed variety of images, even for such an apparently simple abstraction as **MAN**, illustrated in Figure 1.1. A high-level abstraction such as **MAN** has the property that it corresponds to a very large set of possible images, which might be very different from each other from the point of view of simple Euclidean distance in the space of pixel intensities. The set of images for which that label could be appropriate forms a highly convoluted region in pixel space that is not even necessarily a connected region. The **MAN** category can be seen as a high-level abstraction with respect to the space of images. What we call abstraction here can be a category (such as the **MAN** category) or a *feature*, a function of sensory data, which can be discrete (e.g., the input sentence is at the past tense) or continuous (e.g., the input video shows an object moving at 2 meter/second). Many lower-level and intermediate-level concepts (which we also call abstractions here) would be useful to construct a **MAN**-detector. Lower level abstractions are more directly tied to particular percepts, whereas higher level ones are what we call "more abstract" because their connection to actual percepts is more remote, and through other, intermediate-level abstractions.

In addition to the difficulty of coming up with the appropriate intermediate abstractions, the number of visual and semantic categories (such as **MAN**) that we would like an "intelligent" machine to capture is rather large. The focus of deep architecture learning is to automatically discover such abstractions, from the lowest level features to the highest level concepts. Ideally, we would like learning algorithms that enable this discovery with as little human effort as possible, i.e., without having to manually define all necessary abstractions or having to provide a huge set of relevant hand-labeled examples. If these algorithms could tap into the huge resource of text and images on the web, it would certainly help to transfer much of human knowledge into machine-interpretable form.

1.1 How do We Train Deep Architectures?

Deep learning methods aim at learning feature hierarchies with features from higher levels of the hierarchy formed by the composition of lower level features. Automatically learning features at multiple levels of abstraction allow a system to learn complex functions mapping the input to the output directly from data, without depending completely on human-crafted features. This is especially important for higher-level abstractions, which humans often do not know how to specify explicitly in terms of raw sensory input. The ability to automatically learn powerful features will become increasingly important as the amount of data and range of applications to machine learning methods continues to grow.

Depth of architecture refers to the number of levels of composition of non-linear operations in the function learned. Whereas most current learning algorithms correspond to *shallow architectures* (1, 2 or 3 levels), the mammal brain is organized in a *deep architecture* [173] with a given input percept represented at multiple levels of abstraction, each level corresponding to a different area of cortex. Humans often describe such concepts in hierarchical ways, with multiple levels of abstraction. The brain also appears to process information through multiple stages of transformation and representation. This is particularly clear in the primate visual system [173], with its sequence of processing stages: detection of edges, primitive shapes, and moving up to gradually more complex visual shapes.

Inspired by the architectural depth of the brain, neural network researchers had wanted for decades to train deep multi-layer neural networks [19, 191], but no successful attempts were reported before 2006[1]: researchers reported positive experimental results with typically two or three levels (i.e., one or two hidden layers), but training deeper networks consistently yielded poorer results. Something that can be considered a *breakthrough* happened in 2006: Hinton et al. at University of Toronto introduced Deep Belief Networks (DBNs) [73], with a learning algorithm that greedily trains one layer at a time, exploiting an unsupervised learning algorithm for each layer, a Restricted Boltzmann Machine (RBM) [51]. Shortly after, related algorithms based on auto-encoders were proposed [17, 153], apparently exploiting the

[1] Except for neural networks with a special structure called convolutional networks, discussed in Section 4.5.

same principle: *guiding the training of intermediate levels of representation using unsupervised learning, which can be performed locally at each level*. Other algorithms for deep architectures were proposed more recently that exploit neither RBMs nor auto-encoders and that exploit the same principle [131, 202] (see Section 4).

Since 2006, deep networks have been applied with success not only in classification tasks [2, 17, 99, 111, 150, 153, 195], but also in regression [160], dimensionality reduction [74, 158], modeling textures [141], modeling motion [182, 183], object segmentation [114], information retrieval [154, 159, 190], robotics [60], natural language processing [37, 130, 202], and collaborative filtering [162]. Although auto-encoders, RBMs and DBNs can be trained with unlabeled data, in many of the above applications, they have been successfully used to initialize deep *supervised* feedforward neural networks applied to a specific task.

1.2 Intermediate Representations: Sharing Features and Abstractions Across Tasks

Since a deep architecture can be seen as the composition of a series of processing stages, the immediate question that deep architectures raise is: what kind of representation of the data should be found as the output of each stage (i.e., the input of another)? What kind of interface should there be between these stages? A hallmark of recent research on deep architectures is the focus on these intermediate representations: the success of deep architectures belongs to the representations learned in an unsupervised way by RBMs [73], ordinary auto-encoders [17], sparse auto-encoders [150, 153], or denoising auto-encoders [195]. These algorithms (described in more detail in Section 7.2) can be seen as learning to transform one representation (the output of the previous stage) into another, at each step maybe disentangling better the factors of variations underlying the data. As we discuss at length in Section 4, it has been observed again and again that once a good representation has been found at each level, it can be used to initialize and successfully train a deep neural network by supervised gradient-based optimization.

Each level of abstraction found in the brain consists of the "activation" (neural excitation) of a small subset of a large number of features that are, in general, not mutually exclusive. Because these features are not mutually exclusive, they form what is called a *distributed representation* [68, 156]: the information is not localized in a particular neuron but distributed across many. In addition to being distributed, it appears that the brain uses a representation that is *sparse*: only a around 1-4% of the neurons are active together at a given time [5, 113]. Section 3.2 introduces the notion of sparse distributed representation and Section 7.1 describes in more detail the machine learning approaches, some inspired by the observations of the sparse representations in the brain, that have been used to build deep architectures with sparse representations.

Whereas dense distributed representations are one extreme of a spectrum, and sparse representations are in the middle of that spectrum, purely local representations are the other extreme. Locality of representation is intimately connected with the notion of *local generalization*. Many existing machine learning methods are *local in input space*: to obtain a learned function that behaves differently in different regions of data-space, they require different tunable parameters for each of these regions (see more in Section 3.1). Even though statistical efficiency is not necessarily poor when the number of tunable parameters is large, good generalization can be obtained only when adding some form of prior (e.g., that smaller values of the parameters are preferred). When that prior is not task-specific, it is often one that forces the solution to be very smooth, as discussed at the end of Section 3.1. In contrast to learning methods based on local generalization, the total number of patterns that can be distinguished using a distributed representation scales possibly exponentially with the dimension of the representation (i.e., the number of learned features).

In many machine vision systems, learning algorithms have been limited to specific parts of such a processing chain. The rest of the design remains labor-intensive, which might limit the scale of such systems. On the other hand, a hallmark of what we would consider intelligent machines includes a large enough repertoire of concepts. Recognizing **MAN** is not enough. We need algorithms that can tackle a very large set of such tasks and concepts. It seems daunting to manually define that many tasks, and learning becomes essential in this context. Furthermore, it would seem foolish not to exploit the underlying commonalities between these tasks and between the concepts they require. This has been the focus of research on *multi-task learning* [7, 8, 32, 88, 186]. Architectures with multiple levels naturally provide such sharing and re-use of components: the low-level visual features (like edge detectors) and intermediate-level visual features (like object parts) that are useful to detect **MAN** are also useful for a large group of other visual tasks. Deep learning algorithms are based on learning intermediate representations which can be shared across tasks. Hence they can leverage unsupervised data and data from similar tasks [148] to boost performance on large and challenging problems that routinely suffer from a poverty of labelled data, as has been shown by [37], beating the state-of-the-art in several natural language processing tasks. A similar multi-task approach for deep architectures was applied in vision tasks by [2]. Consider a multi-task setting in which there are different outputs for different tasks, all obtained from a shared pool of high-level features. The fact that many of these learned features are shared among m tasks provides sharing of statistical strength in proportion to m. Now consider that these learned high-level features can themselves be represented by combining lower-level intermediate features from a common pool. Again statistical strength can be gained in a similar way, and this strategy can be exploited for every level of a deep architecture.

In addition, learning about a large set of interrelated concepts might provide a key to the kind of broad generalizations that humans appear able to do, which we would not expect from separately trained object detectors, with one detector per visual category. If each high-level category is itself represented through a particular distributed configuration of abstract features from a common pool, generalization to unseen categories could follow naturally from new configurations of these features. Even though only some configurations of these features would present in the training examples, if they represent different aspects of the data, new examples could meaningfully be represented by new configurations of these features.

1.3 Desiderata for Learning AI

Summarizing some of the above issues, and trying to put them in the broader perspective of AI, we put forward a number of requirements we believe to be important for learning algorithms to approach AI, many of which motivate the research are described here:

- Ability to learn complex, highly-varying functions, i.e., with a number of variations much greater than the number of training examples.
- Ability to learn with little human input the low-level, intermediate, and high-level abstractions that would be useful to represent the kind of complex functions needed for AI tasks.
- Ability to learn from a very large set of examples: computation time for training should scale well with the number of examples, i.e., close to linearly.
- Ability to learn from mostly unlabeled data, i.e., to work in the semi-supervised setting, where not all the examples come with complete and correct semantic labels.
- Ability to exploit the synergies present across a large number of tasks, i.e., multi-task learning. These synergies exist because all the AI tasks provide different views on the same underlying reality.
- Strong *unsupervised learning* (i.e., capturing most of the statistical structure in the observed data), which seems essential in the limit of a large number of tasks and when future tasks are not known ahead of time.

Other elements are equally important but are not directly connected to the material in this monograph. They include the ability to learn to represent context of varying length and structure [146], so as to allow machines to operate in a context-dependent stream of observations and produce a stream of actions, the ability to make decisions when actions influence the future observations and future rewards [181], and the ability to influence future observations so as to collect more relevant information about the world, i.e., a form of active learning [34].

1.4 Outline of the Paper

Section 2 reviews theoretical results (which can be skipped without hurting the understanding of the remainder) showing that an architecture with insufficient depth can require many more computational elements, potentially exponentially more (with respect to input size), than architectures whose depth is matched to the task. We claim that insufficient depth can be detrimental for learning. Indeed, if a solution to the task is represented with a very large but shallow architecture (with many computational elements), a lot of training examples might be needed to tune each of these elements and capture a highly varying function. Section 3.1 is also meant to motivate the reader, this time to highlight the limitations of local generalization and local estimation, which we expect to avoid using deep architectures with a distributed representation (Section 3.2).

In later sections, the monograph describes and analyzes some of the algorithms that have been proposed to train deep architectures. Section 4 introduces concepts from the neural networks literature relevant to the task of training deep architectures. We first consider the previous difficulties in training neural networks with many layers, and then introduce unsupervised learning algorithms that could be exploited to initialize deep neural networks. Many of these algorithms (including those for the RBM) are related to the *auto-encoder*: a simple unsupervised algorithm for learning a one-layer model that computes a distributed representation for its input [25, 79, 156]. To fully understand RBMs and many related unsupervised learning algorithms, Section 5 introduces the class of energy-based models, including those used to build generative models with hidden variables such as the Boltzmann Machine. Section 6 focuses on the greedy layer-wise training algorithms for Deep Belief Networks (DBNs) [73] and Stacked Auto-Encoders [17, 153, 195]. Section 7 discusses variants of RBMs and auto-encoders that have been recently proposed to extend and improve them, including the use of sparsity, and the modeling of temporal dependencies. Section 8 discusses algorithms for jointly training all the layers of a Deep Belief Network using variational bounds. Finally, we consider in Section 9 forward looking questions such as the hypothesized difficult optimization problem involved in training deep architectures. In particular, we follow up on the hypothesis that part of the success of current learning strategies for deep architectures is connected to the optimization of lower layers. We discuss the principle of continuation methods, which minimize gradually less smooth versions of the desired cost function, to make a dent in the optimization of deep architectures.

References

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, “A learning algorithm for boltzmann machines,” *Cognitive Science*, vol. 9, pp. 147–169, 1985.

[2] A. Ahmed, K. Yu, W. Xu, Y. Gong, and E. P. Xing, “Training hierarchical feed-forward visual recognition models using transfer learning from pseudo tasks,” in *Proceedings of the 10th European Conference on Computer Vision (ECCV'08)*, pp. 69–82, 2008.

[3] E. L. Allgower and K. Georg, *Numerical Continuation Methods. An Introduction. No. 13 in Springer Series in Computational Mathematics*, Springer-Verlag, 1980.

[4] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan, “An introduction to MCMC for machine learning,” *Machine Learning*, vol. 50, pp. 5–43, 2003.

[5] D. Attwell and S. B. Laughlin, “An energy budget for signaling in the grey matter of the brain,” *Journal of Cerebral Blood Flow And Metabolism*, vol. 21, pp. 1133–1145, 2001.

[6] J. A. Bagnell and D. M. Bradley, “Differentiable sparse coding,” in *Advances in Neural Information Processing Systems 21 (NIPS'08)*, (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), NIPS Foundation, 2009.

[7] J. Baxter, “Learning internal representations,” in *Proceedings of the 8th International Conference on Computational Learning Theory (COLT'95)*, pp. 311–320, Santa Cruz, California: ACM Press, 1995.

[8] J. Baxter, “A Bayesian/information theoretic model of learning via multiple task sampling,” *Machine Learning*, vol. 28, pp. 7–40, 1997.

[9] M. Belkin, I. Matveeva, and P. Niyogi, “Regularization and semi-supervised learning on large graphs,” in *Proceedings of the 17th International Confer-*

ence on Computational Learning Theory (COLT'04), (J. Shawe-Taylor and Y. Singer, eds.), pp. 624–638, Springer, 2004.

[10] M. Belkin and P. Niyogi, "Using manifold structure for partially labeled classification," in *Advances in Neural Information Processing Systems 15 (NIPS'02)*, (S. Becker, S. Thrun, and K. Obermayer, eds.), Cambridge, MA: MIT Press, 2003.

[11] A. J. Bell and T. J. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[12] Y. Bengio and O. Delalleau, "Justifying and generalizing contrastive divergence," *Neural Computation*, vol. 21, no. 6, pp. 1601–1621, 2009.

[13] Y. Bengio, O. Delalleau, and N. Le Roux, "The Curse of highly variable functions for local kernel machines," in *Advances in Neural Information Processing Systems 18 (NIPS'05)*, (Y. Weiss, B. Schölkopf, and J. Platt, eds.), pp. 107–114, Cambridge, MA: MIT Press, 2006.

[14] Y. Bengio, O. Delalleau, and C. Simard, "Decision trees do not generalize to new variations," *Computational Intelligence*, To appear, 2009.

[15] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," in *Advances in Neural Information Processing Systems 13 (NIPS'00)*, (T. Leen, T. Dietterich, and V. Tresp, eds.), pp. 933–938, MIT Press, 2001.

[16] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[17] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems 19 (NIPS'06)*, (B. Schölkopf, J. Platt, and T. Hoffman, eds.), pp. 153–160, MIT Press, 2007.

[18] Y. Bengio, N. Le Roux, P. Vincent, O. Delalleau, and P. Marcotte, "Convex neural networks," in *Advances in Neural Information Processing Systems 18 (NIPS'05)*, (Y. Weiss, B. Schölkopf, and J. Platt, eds.), pp. 123–130, Cambridge, MA: MIT Press, 2006.

[19] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards AI," in *Large Scale Kernel Machines*, (L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, eds.), MIT Press, 2007.

[20] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the Twenty-sixth InternationalConference onMachine Learning (ICML09)*, (L. Bottou and M. Littman, eds.), pp. 41–48, Montreal: ACM, 2009.

[21] Y. Bengio, M. Monperrus, and H. Larochelle, "Non-local estimation of manifold structure," *Neural Computation*, vol. 18, no. 10, pp. 2509–2528, 2006.

[22] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[23] J. Bergstra and Y. Bengio, "Slow, decorrelated features for pretraining complex cell-like networks," in *Advances in Neural Information Processing Systems 22 (NIPS'09)*, (D. Schuurmans, Y. Bengio, C. Williams, J. Lafferty, and A. Culotta, eds.), December 2010.

[24] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152, Pittsburgh: ACM, 1992.

[25] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological Cybernetics*, vol. 59, pp. 291–294, 1988.

[26] M. Brand, "Charting a manifold," in *Advances in Neural Information Processing Systems 15 (NIPS'02)*, (S. Becker, S. Thrun, and K. Obermayer, eds.), pp. 961–968, MIT Press, 2003.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.* Belmont, CA: Wadsworth International Group, 1984.

[29] L. D. Brown, *Fundamentals of Statistical Exponential Families.* 1986. Vol. 9, Inst. of Math. Statist. Lecture Notes Monograph Series.

[30] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 15, no. 12, pp. 4203–4215, 2005.

[31] M. A. Carreira-Perpiñan and G. E. Hinton, "On contrastive divergence learning," in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS'05)*, (R. G. Cowell and Z. Ghahramani, eds.), pp. 33–40, Society for Artificial Intelligence and Statistics, 2005.

[32] R. Caruana, "Multitask connectionist learning," in *Proceedings of the 1993 Connectionist Models Summer School*, pp. 372–379, 1993.

[33] P. Clifford, "Markov random fields in statistics," in *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, (G. Grimmett and D. Welsh, eds.), pp. 19–32, Oxford University Press, 1990.

[34] D. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," in *Advances in Neural Information Processing Systems 7 (NIPS'94)*, (G. Tesauro, D. Touretzky, and T. Leen, eds.), pp. 705–712, Cambridge MA: MIT Press, 1995.

[35] T. F. Coleman and Z. Wu, "Parallel continuation-based global optimization for molecular conformation and protein folding," Technical Report Cornell University, Dept. of Computer Science, 1994.

[36] R. Collobert and S. Bengio, "Links between perceptrons, MLPs and SVMs," in *Proceedings of the Twenty-first International Conference on Machine Learning (ICML'04)*, (C. E. Brodley, ed.), p. 23, New York, NY, USA: ACM, 2004.

[37] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, (W. W. Cohen, A. McCallum, and S. T. Roweis, eds.), pp. 160–167, ACM, 2008.

[38] C. Cortes, P. Haffner, and M. Mohri, "Rational kernels: Theory and algorithms," *Journal of Machine Learning Research*, vol. 5, pp. 1035–1062, 2004.

[39] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[40] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, "On kernel-target alignment," in *Advances in Neural Information Processing Systems 14 (NIPS'01)*, (T. Dietterich, S. Becker, and Z. Ghahramani, eds.), pp. 367–373, 2002.

[41] F. Cucker and D. Grigoriev, "Complexity lower bounds for approximation algebraic computation trees," *Journal of Complexity*, vol. 15, no. 4, pp. 499–512, 1999.

[42] P. Dayan, G. E. Hinton, R. Neal, and R. Zemel, "The Helmholtz machine," *Neural Computation*, vol. 7, pp. 889–904, 1995.

[43] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[44] O. Delalleau, Y. Bengio, and N. L. Roux, "Efficient non-parametric function induction in semi-supervised learning," in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, (R. G. Cowell and Z. Ghahramani, eds.), pp. 96–103, Society for Artificial Intelligence and Statistics, January 2005.

[45] G. Desjardins and Y. Bengio, "Empirical evaluation of convolutional RBMs for vision," Technical Report 1327, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, 2008.

[46] E. Doi, D. C. Balcan, and M. S. Lewicki, "A theoretical analysis of robust coding over noisy overcomplete channels," in *Advances in Neural Information Processing Systems 18 (NIPS'05)*, (Y. Weiss, B. Schölkopf, and J. Platt, eds.), pp. 307–314, Cambridge, MA: MIT Press, 2006.

[47] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[48] S. Duane, A. Kennedy, B. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Phys. Lett. B*, vol. 195, pp. 216–222, 1987.

[49] J. L. Elman, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, pp. 781–799, 1993.

[50] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, "The difficulty of training deep architectures and the effect of unsupervised pre-training," in *Proceedings of The Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS'09)*, pp. 153–160, 2009.

[51] Y. Freund and D. Haussler, "Unsupervised learning of distributions on binary vectors using two layer networks," Technical Report UCSC-CRL-94-25, University of California, Santa Cruz, 1994.

[52] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Machine Learning: Proceedings of Thirteenth International Conference*, pp. 148–156, USA: ACM, 1996.

[53] B. J. Frey, G. E. Hinton, and P. Dayan, "Does the wake-sleep algorithm learn good density estimators?," in *Advances in Neural Information Processing Systems 8 (NIPS'95)*, (D. Touretzky, M. Mozer, and M. Hasselmo, eds.), pp. 661–670, Cambridge, MA: MIT Press, 1996.

[54] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, 1980.

[55] P. Gallinari, Y. LeCun, S. Thiria, and F. Fogelman-Soulie, “Memoires associatives distribuees,” in *Proceedings of COGNITIVA 87*, Paris, La Villette, 1987.

[56] T. Gärtner, “A survey of kernels for structured data,” *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 1, pp. 49–58, 2003.

[57] S. Geman and D. Geman, “Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, November 1984.

[58] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, “Shift-invariant sparse coding for audio classification,” in *Proceedings of the Twenty-third Conference on Uncertainty in Artificial Intelligence (UAI’07)*, 2007.

[59] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR’06)*, pp. 1735–1742, IEEE Press, 2006.

[60] R. Hadsell, A. Erkan, P. Sermanet, M. Scoffier, U. Muller, and Y. LeCun, “Deep belief net learning in a long-range vision system for autonomous off-road driving,” in *Proc. Intelligent Robots and Systems (IROS’08)*, pp. 628–633, 2008.

[61] J. M. Hammersley and P. Clifford, “Markov field on finite graphs and lattices,” Unpublished manuscript, 1971.

[62] J. Håstad, “Almost optimal lower bounds for small depth circuits,” in *Proceedings of the 18th annual ACM Symposium on Theory of Computing*, pp. 6–20, Berkeley, California: ACM Press, 1986.

[63] J. Håstad and M. Goldmann, “On the power of small-depth threshold circuits,” *Computational Complexity*, vol. 1, pp. 113–129, 1991.

[64] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, “The entire regularization path for the support vector machine,” *Journal of Machine Learning Research*, vol. 5, pp. 1391–1415, 2004.

[65] K. A. Heller and Z. Ghahramani, “A nonparametric bayesian approach to modeling overlapping clusters,” in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS’07)*, pp. 187–194, San Juan, Porto Rico: Omnipress, 2007.

[66] K. A. Heller, S. Williamson, and Z. Ghahramani, “Statistical models for partial membership,” in *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML’08)*, (W. W. Cohen, A. McCallum, and S. T. Roweis, eds.), pp. 392–399, ACM, 2008.

[67] G. Hinton and J. Anderson, *Parallel Models of Associative Memory*. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1981.

[68] G. E. Hinton, “Learning distributed representations of concepts,” in *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pp. 1–12, Amherst: Lawrence Erlbaum, Hillsdale, 1986.

[69] G. E. Hinton, “Products of experts,” in *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN)*, vol. 1, pp. 1–6, Edinburgh, Scotland: IEE, 1999.

[70] G. E. Hinton, “Training products of experts by minimizing contrastive divergence,” *Neural Computation*, vol. 14, pp. 1771–1800, 2002.

[71] G. E. Hinton, "To recognize shapes, first learn to generate images," Technical Report UTML TR 2006-003, University of Toronto, 2006.

[72] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, "The wake-sleep algorithm for unsupervised neural networks," *Science*, vol. 268, pp. 1558–1161, 1995.

[73] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[74] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[75] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[76] G. E. Hinton and T. J. Sejnowski, "Learning and relearning in Boltzmann machines," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, (D. E. Rumelhart and J. L. McClelland, eds.), pp. 282–317, Cambridge, MA: MIT Press, 1986.

[77] G. E. Hinton, T. J. Sejnowski, and D. H. Ackley, "Boltzmann machines: Constraint satisfaction networks that learn," Technical Report TR-CMU-CS-84-119, Carnegie-Mellon University, Dept. of Computer Science, 1984.

[78] G. E. Hinton, M. Welling, Y. W. Teh, and S. Osindero, "A new view of ICA," in *Proceedings of 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA'01)*, pp. 746–751, San Diego, CA, 2001.

[79] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length, and Helmholtz free energy," in *Advances in Neural Information Processing Systems 6 (NIPS'93)*, (D. Cowan, G. Tesauro, and J. Alspector, eds.), pp. 3–10, Morgan Kaufmann Publishers, Inc., 1994.

[80] T. K. Ho, "Random decision forest," in *3rd International Conference on Document Analysis and Recognition (ICDAR'95)*, pp. 278–282, Montreal, Canada, 1995.

[81] S. Hochreiter Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München, 1991.

[82] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 498–520, 1933.

[83] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex," *Journal of Physiology (London)*, vol. 160, pp. 106–154, 1962.

[84] A. Hyvärinen, "Estimation of non-normalized statistical models using score matching," *Journal of Machine Learning Research*, vol. 6, pp. 695–709, 2005.

[85] A. Hyvärinen, "Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables," *IEEE Transactions on Neural Networks*, vol. 18, pp. 1529–1531, 2007.

[86] A. Hyvärinen, "Some extensions of score matching," *Computational Statistics and Data Analysis*, vol. 51, pp. 2499–2512, 2007.

[87] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley-Interscience, May 2001.

[88] N. Intrator and S. Edelman, "How to make a low-dimensional representation suitable for diverse tasks," *Connection Science, Special issue on Transfer in Neural Networks*, vol. 8, pp. 205–224, 1996.

[89] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," Available from http://www.cse.ucsc.edu/ haussler/pubs.html, Preprint, Dept.of Computer Science, Univ. of California. A shorter version is in Advances in Neural Information Processing Systems 11, 1998.

[90] N. Japkowicz, S. J. Hanson, and M. A. Gluck, "Nonlinear autoassociation is not equivalent to PCA," *Neural Computation*, vol. 12, no. 3, pp. 531–545, 2000.

[91] M. I. Jordan, *Learning in Graphical Models*. Dordrecht, Netherlands: Kluwer, 1998.

[92] K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "Fast inference in sparse coding algorithms with applications to object recognition," Technical Report, Computational and Biological Learning Lab, Courant Institute, NYU, Technical Report CBLL-TR-2008-12-01, 2008.

[93] S. Kirkpatrick, C. D. G. Jr., and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, 1983.

[94] U. Köster and A. Hyvärinen, "A two-layer ICA-like model estimated by score matching," in *Int. Conf. Artificial Neural Networks (ICANN'2007)*, pp. 798–807, 2007.

[95] K. A. Krueger and P. Dayan, "Flexible shaping: How learning in small steps helps," *Cognition*, vol. 110, pp. 380–394, 2009.

[96] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Gahoui, and M. Jordan, "Learning the kernel matrix with semi-definite programming," in *Proceedings of the Nineteenth International Conference on Machine Learning (ICML'02)*, (C. Sammut and A. G. Hoffmann, eds.), pp. 323–330, Morgan Kaufmann, 2002.

[97] H. Larochelle and Y. Bengio, "Classification using discriminative restricted Boltzmann machines," in *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, (W. W. Cohen, A. McCallum, and S. T. Roweis, eds.), pp. 536–543, ACM, 2008.

[98] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 10, pp. 1–40, 2009.

[99] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML'07)*, (Z. Ghahramani, ed.), pp. 473–480, ACM, 2007.

[100] J. A. Lasserre, C. M. Bishop, and T. P. Minka, "Principled hybrids of generative and discriminative models," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'06)*, pp. 87–94, Washington, DC, USA, 2006. IEEE Computer Society.

[101] Y. Le Cun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[102] N. Le Roux and Y. Bengio, "Representational power of restricted boltzmann machines and deep belief networks," *Neural Computation*, vol. 20, no. 6, pp. 1631–1649, 2008.

[103] Y. LeCun, "Modèles connexionistes de l'apprentissage," PhD thesis, Université de Paris VI, 1987.

[104] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[105] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient BackProp," in *Neural Networks: Tricks of the Trade*, (G. B. Orr and K.-R. Müller, eds.), pp. 9–50, Springer, 1998.

[106] Y. LeCun, S. Chopra, R. M. Hadsell, M.-A. Ranzato, and F.-J. Huang, "A tutorial on energy-based learning," in *Predicting Structured Data*, pp. 191–246, G. Bakir and T. Hofman and B. Scholkopf and A. Smola and B. Taskar: MIT Press, 2006.

[107] Y. LeCun and F. Huang, "Loss functions for discriminative training of energy-based models," in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS'05)*, (R. G. Cowell and Z. Ghahramani, eds.), 2005.

[108] Y. LeCun, F.-J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'04),* vol. 2, pp. 97–104, Los Alamitos, CA, USA: IEEE Computer Society, 2004.

[109] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems 19 (NIPS'06)*, (B. Schölkopf, J. Platt, and T. Hoffman, eds.), pp. 801–808, MIT Press, 2007.

[110] H. Lee, C. Ekanadham, and A. Ng, "Sparse deep belief net model for visual area V2," in *Advances in Neural Information Processing Systems 20 (NIPS'07)*, (J. Platt, D. Koller, Y. Singer, and S. P. Roweis, eds.), Cambridge, MA: MIT Press, 2008.

[111] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the Twenty-sixth International Conference on Machine Learning (ICML'09)*, (L. Bottou and M. Littman, eds.), Montreal (Qc), Canada: ACM, 2009.

[112] T.-S. Lee and D. Mumford, "Hierarchical bayesian inference in the visual cortex," *Journal of Optical Society of America, A*, vol. 20, no. 7, pp. 1434–1448, 2003.

[113] P. Lennie, "The cost of cortical computation," *Current Biology*, vol. 13, pp. 493–497, Mar 18 2003.

[114] I. Levner, *Data Driven Object Segmentation.* 2008. PhD thesis, Department of Computer Science, University of Alberta.

[115] M. Lewicki and T. Sejnowski, "Learning nonlinear overcomplete representations for efficient coding," in *Advances in Neural Information Processing Systems 10 (NIPS'97)*, (M. Jordan, M. Kearns, and S. Solla, eds.), pp. 556–562, Cambridge, MA, USA: MIT Press, 1998.

[116] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[117] M. Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*. New York, NY: Springer, Second ed., 1997.

[118] P. Liang and M. I. Jordan, "An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators," in *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, (W. W. Cohen, A. McCallum, and S. T. Roweis, eds.), pp. 584–591, New York, NY, USA: ACM, 2008.

[119] T. Lin, B. G. Horne, P. Tino, and C. L. Giles, "Learning long-term dependencies is not as difficult with NARX recurrent neural networks," Technical Report UMICAS-TR-95-78, Institute for Advanced Computer Studies, University of Mariland, 1995.

[120] G. Loosli, S. Canu, and L. Bottou, "Training invariant support vector machines using selective sampling," in *Large Scale Kernel Machines*, (L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, eds.), pp. 301–320, Cambridge, MA: MIT Press, 2007.

[121] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Advances in Neural Information Processing Systems 21 (NIPS'08)*, (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 1033–1040, 2009. NIPS Foundation.

[122] J. L. McClelland and D. E. Rumelhart, "An interactive activation model of context effects in letter perception," *Psychological Review*, pp. 375–407, 1981.

[123] J. L. McClelland and D. E. Rumelhart, *Explorations in parallel distributed processing*. Cambridge: MIT Press, 1988.

[124] J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 2. Cambridge: MIT Press, 1986.

[125] W. S. McCulloch and W. Pitts, "A logical calculus of ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943.

[126] R. Memisevic and G. E. Hinton, "Unsupervised learning of image transformations," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'07)*, 2007.

[127] E. Mendelson, *Introduction to Mathematical Logic, 4th ed.* 1997. Chapman & Hall.

[128] R. Miikkulainen and M. G. Dyer, "Natural language processing with modular PDP networks and distributed lexicon," *Cognitive Science*, vol. 15, pp. 343–399, 1991.

[129] A. Mnih and G. E. Hinton, "Three new graphical models for statistical language modelling," in *Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML'07)*, (Z. Ghahramani, ed.), pp. 641–648, ACM, 2007.

[130] A. Mnih and G. E. Hinton, “A scalable hierarchical distributed language model,” in *Advances in Neural Information Processing Systems 21 (NIPS’08)*, (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 1081–1088, 2009.

[131] H. Mobahi, R. Collobert, and J. Weston, “Deep Learning from temporal coherence in video,” in *Proceedings of the 26th International Conference on Machine Learning*, (L. Bottou and M. Littman, eds.), pp. 737–744, Montreal: Omnipress, June 2009.

[132] J. More and Z. Wu, “Smoothing techniques for macromolecular global optimization,” in *Nonlinear Optimization and Applications*, (G. D. Pillo and F. Giannessi, eds.), Plenum Press, 1996.

[133] I. Murray and R. Salakhutdinov, “Evaluating probabilities under high-dimensional latent variable models,” in *Advances in Neural Information Processing Systems 21 (NIPS’08),* vol. 21, (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 1137–1144, 2009.

[134] J. Mutch and D. G. Lowe, “Object class recognition and localization using sparse features with limited receptive fields,” *International Journal of Computer Vision*, vol. 80, no. 1, pp. 45–57, 2008.

[135] R. M. Neal, “Connectionist learning of belief networks,” *Artificial Intelligence*, vol. 56, pp. 71–113, 1992.

[136] R. M. Neal, “Bayesian learning for neural networks,” PhD thesis, Department of Computer Science, University of Toronto, 1994.

[137] A. Y. Ng and M. I. Jordan, “On Discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes,” in *Advances in Neural Information Processing Systems 14 (NIPS’01)*, (T. Dietterich, S. Becker, and Z. Ghahramani, eds.), pp. 841–848, 2002.

[138] J. Niebles and L. Fei-Fei, “A hierarchical model of shape and appearance for human action classification,” in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR’07)*, 2007.

[139] B. A. Olshausen and D. J. Field, “Sparse coding with an overcomplete basis set: a strategy employed by V1?,” *Vision Research*, vol. 37, pp. 3311–3325, December 1997.

[140] P. Orponen, “Computational complexity of neural networks: a survey,” *Nordic Journal of Computing*, vol. 1, no. 1, pp. 94–110, 1994.

[141] S. Osindero and G. E. Hinton, “Modeling image patches with a directed hierarchy of Markov random field,” in *Advances in Neural Information Processing Systems 20 (NIPS’07)*, (J. Platt, D. Koller, Y. Singer, and S. Roweis, eds.), pp. 1121–1128, Cambridge, MA: MIT Press, 2008.

[142] B. Pearlmutter and L. C. Parra, “A context-sensitive generalization of ICA,” in *International Conference On Neural Information Processing*, (L. Xu, ed.), pp. 151–157, Hong-Kong, 1996.

[143] E. Pérez and L. A. Rendell, “Learning despite concept variation by finding structure in attribute-based data,” in *Proceedings of the Thirteenth International Conference on Machine Learning (ICML’96)*, (L. Saitta, ed.), pp. 391–399, Morgan Kaufmann, 1996.

[144] G. B. Peterson, "A day of great illumination: B. F. Skinner's discovery of shaping," *Journal of the Experimental Analysis of Behavior*, vol. 82, no. 3, pp. 317–328, 2004.

[145] N. Pinto, J. DiCarlo, and D. Cox, "Establishing good benchmarks and baselines for face recognition," in *ECCV 2008 Faces in 'Real-Life' Images Workshop*, 2008. Marseille France, Erik Learned-Miller and Andras Ferencz and Frédéric Jurie.

[146] J. B. Pollack, "Recursive distributed representations," *Artificial Intelligence*, vol. 46, no. 1, pp. 77–105, 1990.

[147] L. R. Rabiner and B. H. Juang, "An Introduction to hidden Markov models," *IEEE ASSP Magazine*, pp. 257–285, january 1986.

[148] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML'07)*, (Z. Ghahramani, ed.), pp. 759–766, ACM, 2007.

[149] M. Ranzato, Y. Boureau, S. Chopra, and Y. LeCun, "A unified energy-based framework for unsupervised learning," in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS'07)*, San Juan, Porto Rico: Omnipress, 2007.

[150] M. Ranzato, Y.-L. Boureau, and Y. LeCun, "Sparse feature learning for deep belief networks," in *Advances in Neural Information Processing Systems 20 (NIPS'07)*, (J. Platt, D. Koller, Y. Singer, and S. Roweis, eds.), pp. 1185–1192, Cambridge, MA: MIT Press, 2008.

[151] M. Ranzato, F. Huang, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'07)*, IEEE Press, 2007.

[152] M. Ranzato and Y. LeCun, "A sparse and locally shift invariant feature extractor applied to document images," in *International Conference on Document Analysis and Recognition (ICDAR'07)*, pp. 1213–1217, Washington, DC, USA: IEEE Computer Society, 2007.

[153] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," in *Advances in Neural Information Processing Systems 19 (NIPS'06)*, (B. Schölkopf, J. Platt, and T. Hoffman, eds.), pp. 1137–1144, MIT Press, 2007.

[154] M. Ranzato and M. Szummer, "Semi-supervised learning of compact document representations with deep networks," in *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08),* vol. 307, (W. W. Cohen, A. McCallum, and S. T. Roweis, eds.), pp. 792–799, ACM, 2008.

[155] S. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[156] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[157] D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Vol. 1, Cambridge: MIT Press, 1986.

[158] R. Salakhutdinov and G. E. Hinton, “Learning a nonlinear embedding by preserving class neighbourhood structure,” in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS’07)*, San Juan, Porto Rico: Omnipress, 2007.

[159] R. Salakhutdinov and G. E. Hinton, “Semantic hashing,” in *Proceedings of the 2007 Workshop on Information Retrieval and applications of Graphical Models (SIGIR 2007)*, Amsterdam: Elsevier, 2007.

[160] R. Salakhutdinov and G. E. Hinton, “Using deep belief nets to learn covariance kernels for Gaussian processes,” in *Advances in Neural Information Processing Systems 20 (NIPS’07)*, (J. Platt, D. Koller, Y. Singer, and S. Roweis, eds.), pp. 1249–1256, Cambridge, MA: MIT Press, 2008.

[161] R. Salakhutdinov and G. E. Hinton, “Deep Boltzmann machines,” in *Proceedings of The Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS’09)*, vol. 5, pp. 448–455, 2009.

[162] R. Salakhutdinov, A. Mnih, and G. E. Hinton, “Restricted Boltzmann machines for collaborative filtering,” in *Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML’07)*, (Z. Ghahramani, ed.), pp. 791–798, New York, NY, USA: ACM, 2007.

[163] R. Salakhutdinov and I. Murray, “On the quantitative analysis of deep belief networks,” in *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML’08)*, (W. W. Cohen, A. McCallum, and S. T. Roweis, eds.), pp. 872–879, ACM, 2008.

[164] L. K. Saul, T. Jaakkola, and M. I. Jordan, “Mean field theory for sigmoid belief networks,” *Journal of Artificial Intelligence Research*, vol. 4, pp. 61–76, 1996.

[165] M. Schmitt, “Descartes’ rule of signs for radial basis function neural networks,” *Neural Computation*, vol. 14, no. 12, pp. 2997–3011, 2002.

[166] B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in Kernel Methods — Support Vector Learning*. Cambridge, MA: MIT Press, 1999.

[167] B. Schölkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. Smola, “Input space versus feature space in kernel-based methods,” *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.

[168] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[169] H. Schwenk, “Efficient training of large neural networks for language modeling,” in *International Joint Conference on Neural Networks (IJCNN)*, pp. 3050–3064, 2004.

[170] H. Schwenk and J.-L. Gauvain, “Connectionist language modeling for large vocabulary continuous speech recognition,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 765–768, Orlando, Florida, 2002.

[171] H. Schwenk and J.-L. Gauvain, “Building continuous space language models for transcribing European languages,” in *Interspeech*, pp. 737–740, 2005.

[172] H. Schwenk and M. Milgram, “Transformation invariant autoassociation with application to handwritten character recognition,” in *Advances in Neural*

Information Processing Systems 7 (NIPS'94), (G. Tesauro, D. Touretzky, and T. Leen, eds.), pp. 991–998, MIT Press, 1995.

[173] T. Serre, G. Kreiman, M. Kouh, C. Cadieu, U. Knoblich, and T. Poggio, "A quantitative theory of immediate visual recognition," *Progress in Brain Research, Computational Neuroscience: Theoretical Insights into Brain Function*, vol. 165, pp. 33–56, 2007.

[174] S. H. Seung, "Learning continuous attractors in recurrent networks," in *Advances in Neural Information Processing Systems 10 (NIPS'97)*, (M. Jordan, M. Kearns, and S. Solla, eds.), pp. 654–660, MIT Press, 1998.

[175] D. Simard, P. Y. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks," in *International Conference on Document Analysis and Recognition (ICDAR'03)*, p. 958, Washington, DC, USA: IEEE Computer Society, 2003.

[176] P. Y. Simard, Y. LeCun, and J. Denker, "Efficient pattern recognition using a new transformation distance," in *Advances in Neural Information Processing Systems 5 (NIPS'92)*, (C. Giles, S. Hanson, and J. Cowan, eds.), pp. 50–58, Morgan Kaufmann, San Mateo, 1993.

[177] B. F. Skinner, "Reinforcement today," *American Psychologist*, vol. 13, pp. 94–99, 1958.

[178] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in *Parallel Distributed Processing*, vol. 1, (D. E. Rumelhart and J. L. McClelland, eds.), pp. 194–281, Cambridge: MIT Press, 1986. ch. 6.

[179] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Describing visual scenes using transformed objects and parts," *International Journal of Computer Vision*, vol. 77, pp. 291–330, 2007.

[180] I. Sutskever and G. E. Hinton, "Learning multilevel distributed representations for high-dimensional sequences," in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS'07)*, San Juan, Porto Rico: Omnipress, 2007.

[181] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction.* MIT Press, 1998.

[182] G. Taylor and G. Hinton, "Factored conditional restricted Boltzmann machines for modeling motion style," in *Proceedings of the 26th International Conference on Machine Learning (ICML'09)*, (L. Bottou and M. Littman, eds.), pp. 1025–1032, Montreal: Omnipress, June 2009.

[183] G. Taylor, G. E. Hinton, and S. Roweis, "Modeling human motion using binary latent variables," in *Advances in Neural Information Processing Systems 19 (NIPS'06)*, (B. Schölkopf, J. Platt, and T. Hoffman, eds.), pp. 1345–1352, Cambridge, MA: MIT Press, 2007.

[184] Y. Teh, M. Welling, S. Osindero, and G. E. Hinton, "Energy-based models for sparse overcomplete representations," *Journal of Machine Learning Research*, vol. 4, pp. 1235–1260, 2003.

[185] J. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[186] S. Thrun, "Is learning the n-th thing any easier than learning the first?," in *Advances in Neural Information Processing Systems 8 (NIPS'95)*, (D. Touretzky, M. Mozer, and M. Hasselmo, eds.), pp. 640–646, Cambridge, MA: MIT Press, 1996.

[187] T. Tieleman, "Training restricted Boltzmann machines using approximations to the likelihood gradient," in *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, (W. W. Cohen, A. McCallum, and S. T. Roweis, eds.), pp. 1064–1071, ACM, 2008.

[188] T. Tieleman and G. Hinton, "Using fast weights to improve persistent contrastive divergence," in *Proceedings of the Twenty-sixth International Conference on Machine Learning (ICML'09)*, (L. Bottou and M. Littman, eds.), pp. 1033–1040, New York, NY, USA: ACM, 2009.

[189] I. Titov and J. Henderson, "Constituent parsing with incremental sigmoid belief networks," in *Proc. 45th Meeting of Association for Computational Linguistics (ACL'07)*, pp. 632–639, Prague, Czech Republic, 2007.

[190] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large databases for recognition," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'08)*, pp. 1–8, 2008.

[191] P. E. Utgoff and D. J. Stracuzzi, "Many-layered learning," *Neural Computation*, vol. 14, pp. 2497–2539, 2002.

[192] L. van der Maaten and G. E. Hinton, "Visualizing data using t-Sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, November 2008.

[193] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

[194] R. Vilalta, G. Blix, and L. Rendell, "Global data analysis and the fragmentation problem in decision tree induction," in *Proceedings of the 9th European Conference on Machine Learning (ECML'97)*, pp. 312–327, Springer-Verlag, 1997.

[195] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, (W. W. Cohen, A. McCallum, and S. T. Roweis, eds.), pp. 1096–1103, ACM, 2008.

[196] L. Wang and K. L. Chan, "Learning kernel parameters by using class separability measure," 6th kernel machines workshop, in conjunction with Neural Information Processing Systems (NIPS), 2002.

[197] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *Proc. 6th Europ. Conf. Comp. Vis., ECCV2000*, pp. 18–32, Dublin, 2000.

[198] I. Wegener, *The Complexity of Boolean Functions*. John Wiley & Sons, 1987.

[199] Y. Weiss, "Segmentation using eigenvectors: a unifying view," in *Proceedings IEEE International Conference on Computer Vision (ICCV'99)*, pp. 975–982, 1999.

[200] M. Welling, M. Rosen-Zvi, and G. E. Hinton, "Exponential family harmoniums with an application to information retrieval," in *Advances in Neural Information Processing Systems 17 (NIPS'04)*, (L. Saul, Y. Weiss, and L. Bottou, eds.), pp. 1481–1488, Cambridge, MA: MIT Press, 2005.

[201] M. Welling, R. Zemel, and G. E. Hinton, "Self-supervised boosting," in *Advances in Neural Information Processing Systems 15 (NIPS'02)*, (S. Becker, S. Thrun, and K. Obermayer, eds.), pp. 665–672, MIT Press, 2003.

[202] J. Weston, F. Ratle, and R. Collobert, "Deep learning via semi-supervised embedding," in *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, (W. W. Cohen, A. McCallum, and S. T. Roweis, eds.), pp. 1168–1175, New York, NY, USA: ACM, 2008.

[203] C. K. I. Williams and C. E. Rasmussen, "Gaussian processes for regression," in *Advances in neural information processing systems 8 (NIPS'95)*, (D. Touretzky, M. Mozer, and M. Hasselmo, eds.), pp. 514–520, Cambridge, MA: MIT Press, 1996.

[204] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Computation*, vol. 14, no. 4, pp. 715–770, 2002.

[205] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–249, 1992.

[206] Z. Wu, "Global continuation for distance geometry problems," *SIAM Journal of Optimization*, vol. 7, pp. 814–836, 1997.

[207] P. Xu, A. Emami, and F. Jelinek, "Training connectionist models for the structured language model," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP'2003),* vol. 10, pp. 160–167, 2003.

[208] A. Yao, "Separating the polynomial-time hierarchy by oracles," in *Proceedings of the 26th Annual IEEE Symposium on Foundations of Computer Science*, pp. 1–10, 1985.

[209] D. Zhou, O. Bousquet, T. Navin Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems 16 (NIPS'03)*, (S. Thrun, L. Saul, and B. Schölkopf, eds.), pp. 321–328, Cambridge, MA: MIT Press, 2004.

[210] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proceedings of the Twenty International Conference on Machine Learning (ICML'03)*, (T. Fawcett and N. Mishra, eds.), pp. 912–919, AAAI Press, 2003.

[211] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proceedings of the Twenty International Conference on Machine Learning (ICML'03)*, (T. Fawcett and N. Mishra, eds.), pp. 928–936, AAAI Press, 2003.