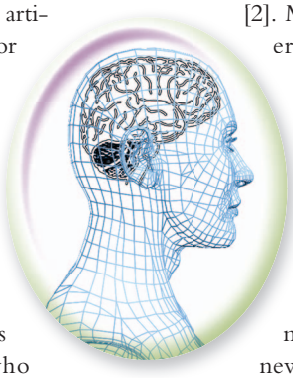
Itamar Arel, Derek C. Rose,
and Thomas P. Karnowski
The University of Tennessee, USA

Research Frontier

Deep Machine Learning—A New Frontier in Artificial Intelligence Research

I. Introduction

Mimicking the efficiency and robustness by which the human brain represents information has been a core challenge in artificial intelligence research for decades. Humans are exposed to myriad of sensory data received every second of the day and are somehow able to capture critical aspects of this data in a way that allows for its future use in a concise manner. Over 50 years ago, Richard Bellman, who introduced dynamic programming theory and pioneered the field of optimal control, asserted that high dimensionality of data is a fundamental hurdle in many science and engineering applications. The main difficulty that arises, particularly in the context of pattern classification applications, is that the learning complexity grows exponentially with linear increase in the dimensionality of the data. He coined this phenomenon the curse of dimensionality [1]. The mainstream approach of overcoming "the curse" has been to pre-process the data in a manner that would reduce its dimensionality to that which can be effectively processed, for example by a classification engine. This dimensionality reduction scheme is often referred to as feature extraction. As a result, it can be argued that the intelligence behind many pattern recognition systems has shifted to the human-engineered feature extraction process, which at times can be challenging and highly application-dependent [2]. Moreover, if incomplete or erroneous features are extracted, the classification process is inherently limited in performance.



© BRAND X PICTURES

Recent neuroscience findings have provided insight into the principles governing information representation in the mammalian brain, leading to new ideas for designing systems that represent information. One of the key findings has been that the neocortex, which is associated with many cognitive abilities, does not explicitly pre-process sensory signals, but rather allows them to propagate through a complex hierarchy [3] of modules that, over time, learn to represent observations based on the regularities they exhibit [4]. This discovery motivated the emergence of the subfield of deep machine learning, which focuses on computational models for information representation that exhibit similar characteristics to that of the neocortex.

In addition to the spatial dimensionality of real-life data, the temporal component also plays a key role. An observed sequence of patterns often conveys a meaning to the observer, whereby independent fragments of this sequence would be hard to decipher in isolation. Meaning is often inferred from events or observations that are received closely in time [5] [6]. To that end, modeling the temporal component of the observations plays a critical role in effective information representation. Capturing spatiotemporal dependencies, based on regularities in the observations, is therefore viewed as a fundamental goal for deep learning systems.

Assuming robust deep learning is achieved, it would be possible to train such a hierarchical network on a large set of observations and later extract signals from this network to a relatively simple classification engine for the purpose of robust pattern recognition. Robustness here refers to the ability to exhibit classification invariance to a diverse range of transformations and distortions, including noise, scale, rotation, various lighting conditions, displacement, etc.

This article provides an overview of the mainstream deep learning approaches and research directions proposed over the past decade. It is important to emphasize that each approach has strengths and weaknesses, depending on the application and context in which it is being used. Thus, this article presents a summary on the current state of the deep machine learning field and some perspective into how it may evolve. Convolutional Neural Networks (CNNs) and Deep Belief Networks (DBNs) (and their respective variations) are focused on primarily because they are well established in the deep learning field and show great promise for future work. Section II introduces CNNs and subsequently follows by details of DBNs in Section III. For an excellent further

Digital Object Identifier 10.1109/MCI.2010.938364

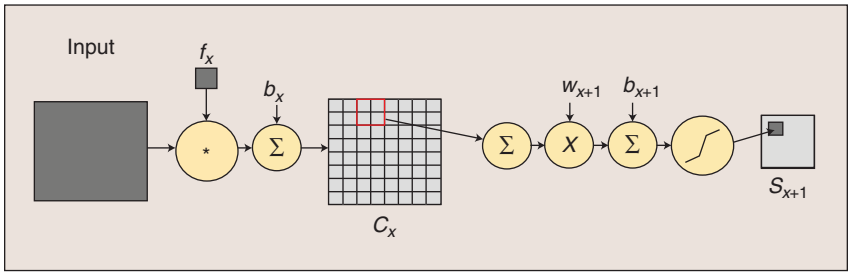


FIGURE 1 The convolution and subsampling process: the convolution process consists of convolving an input (image for the first stage or feature map for later stages) with a trainable filter f_x then adding a trainable bias b_x to produce the convolution layer C_x. The subsampling consists of summing a neighborhood (four pixels), weighting by scalar w_{x+1}, adding trainable bias b_{x+1}, and passing through a sigmoid function to produce a roughly 2x smaller feature map S_{x+1}.

in-depth look at the foundations of these technologies, the reader is referred to [7]. Section IV contains other deep architectures that are currently being proposed. Section V contains a brief note about how this research has impacted government and industry initiatives. The conclusion provides a perspective of the potential impact of deep-layered architectures as well as key questions that remain to be answered.

II. Convolutional Neural Networks

CNNs [8] [9] are a family of multi-layer neural networks particularly designed for use on two-dimensional data, such as images and videos. CNNs are influenced by earlier work in time-delay neural networks (TDNN), which reduce learning computation requirements by sharing weights in a temporal dimension and are intended for speech and time-series processing [53]. CNNs are the first truly successful deep learning approach where many layers of a hierarchy are successfully trained in a robust manner. A CNN is a choice of topology or architecture that leverages spatial relationships to reduce the number of parameters which must be learned and thus improves upon general feed-forward back propagation training. CNNs were proposed as a deep learning framework that is motivated by minimal data preprocessing requirements. In CNNs, small portions of the image (dubbed a local receptive field) are treated as inputs to the lowest layer of the hierarchical structure. Information generally propagates through the different layers of the network whereby at each layer digital filtering is applied in order to obtain salient features of the data observed. The method provides a level of invariance to shift, scale and rotation as the local receptive field allows the neuron or processing unit access to elementary features such as oriented edges or corners.

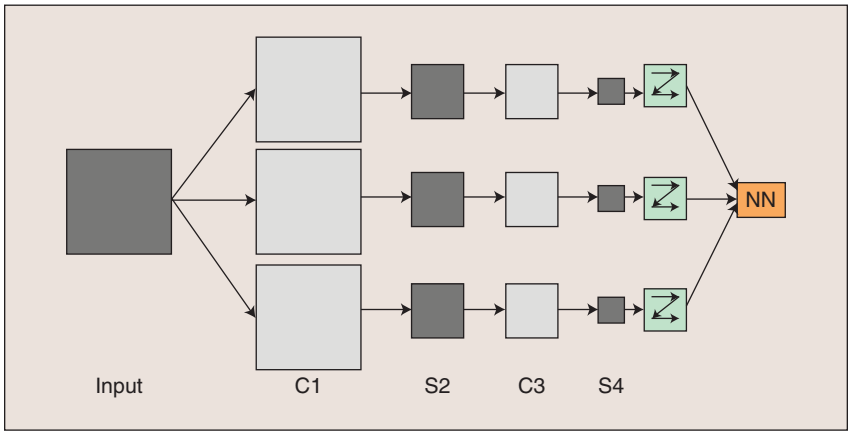


FIGURE 2 Conceptual example of convolutional neural network. The input image is convolved with three trainable filters and biases as in Figure 1 to produce three feature maps at the C1 level. Each group of four pixels in the feature maps are added, weighted, combined with a bias, and passed through a sigmoid function to produce the three feature maps at S2. These are again filtered to produce the C3 level. The hierarchy then produces S4 in a manner analogous to S2. Finally these pixel values are rasterized and presented as a single vector input to the "conventional" neural network at the output.

One of the seminal papers on the topic [8] describes an application of CNNs to the classification handwritten digits in the MNIST database. Essentially, the input image is convolved with a set of N small filters whose coefficients are either trained or pre-determined using some criteria. Thus, the first (or lowest) layer of the network consists of "feature maps" which are the result of the convolution processes, with an additive bias and possibly a compression or normalization of the features. This initial stage is followed by a subsampling (typically a 2 $\times$ 2 averaging operation) that further reduces the dimensionality and offers some robustness to spatial shifts (see Figure 1). The subsampled feature map then receives a weighting and trainable bias and finally propagates through an activation function. Some variants of this exist with as few as one map per layer [13] or summations of multiple maps [8].

When the weighting is small, the activation function is nearly linear and the result is a blurring of the image; other weightings can cause the activation output to resemble an AND or OR function. These outputs form a new feature map that is then passed through another sequence of convolution, sub-sampling and activation function flow, as illustrated in Figure 2. This process can be repeated an arbitrary number of times. It should be noted that subsequent layers can combine one or more of the previous layers; for example, in [8] the initial six feature maps are combined to form 16 feature maps in the subsequent layer. As described in [33], CNNs create their invariance to object translations by a method dubbed "feature pooling" (the S layers in Figure 2). However, feature pooling is hand crafted by the network organizer, not trained or learned by the system; in CNNs, the pooling is "tuned" by parameters in the learning process but the basic mechanism (the

combination of inputs to the S layers, for example) are set by the network designer. Finally, at the final stage of the process, the activation outputs are forwarded to a conventional feedforward neural network that produces the final output of the system.

The intimate relationship between the layers and spatial information in CNNs renders them well suited for image processing and understanding, and they generally perform well at autonomously extracting salient features from images. In some cases Gabor filters have been used as an initial pre-processing step to emulate the human visual response to visual excitation [10]. In more recent work, researchers have applied CNNs to various machine learning problems including face detection [11] [13], document analysis [38], and speech detection [12]. CNNs have recently [25] been trained with a temporal coherence objective to leverage the frame-to-frame coherence found in videos, though this objective need not be specific to CNNs.

III. Deep Belief Networks

DBNs, initially introduced in [14], are probabilistic generative models that stand in contrast to the discriminative nature of traditional neural nets. Generative models provide a joint probability distribution over observable data and labels, facilitating the estimation of both $P(Observation|Label)$ as well as $P(Label|Observation)$, while discriminative models are limited to the latter, $P(Label|Observation)$. DBNs address problems encountered when traditionally applying back-propagation to deeply-layered neural networks, namely: (1) necessity of a substantial labeled data set for training, (2) slow learning (i.e. convergence) times, and (3) inadequate parameter selection techniques that lead to poor local optima.

DBNs are composed of several layers of Restricted Boltzmann Machines, a type of neural network (see Figure 3). These networks are "restricted" to a single visible layer and single hidden layer, where connections are formed between the layers (units within a layer are not connected). The hidden units are trained to capture higher-order data correlations that are observed at the visible units. Initially, aside from the top two layers, which form an associative memory, the layers of a DBN are connected only by directed top-down generative weights. RBMs are attractive as a building block, over more traditional and deeply layered sigmoid belief networks, due to their ease of learning these connection weights. To obtain generative weights, the initial pre-training occurs in an unsupervised greedy layer-by-layer manner, enabled by what Hinton has termed *contrastive divergence* [15]. During this training phase, a vector $\overline{v}$ is presented to the visible units that forward values to the hidden units. Going in reverse, the visible unit inputs are then stochastically found in an attempt to reconstruct the original input. Finally, these new visible neuron activations are forwarded such that one step reconstruction hidden unit activations, $\overline{h}$, can be attained. Performing these back and forth steps is a process known as Gibbs sampling, and the difference in the correlation of the hidden activations and visible inputs forms the basis for a weight update. Training time is significantly reduced as it can be shown that only a single step is needed to approximate maximum likelihood learning. Each layer added to the network improves the log-probability of the training data, which we can think of as increasing true representational power. This meaningful expansion, in conjunction with the utilization of unlabeled data, is a critical component in any deep learning application.

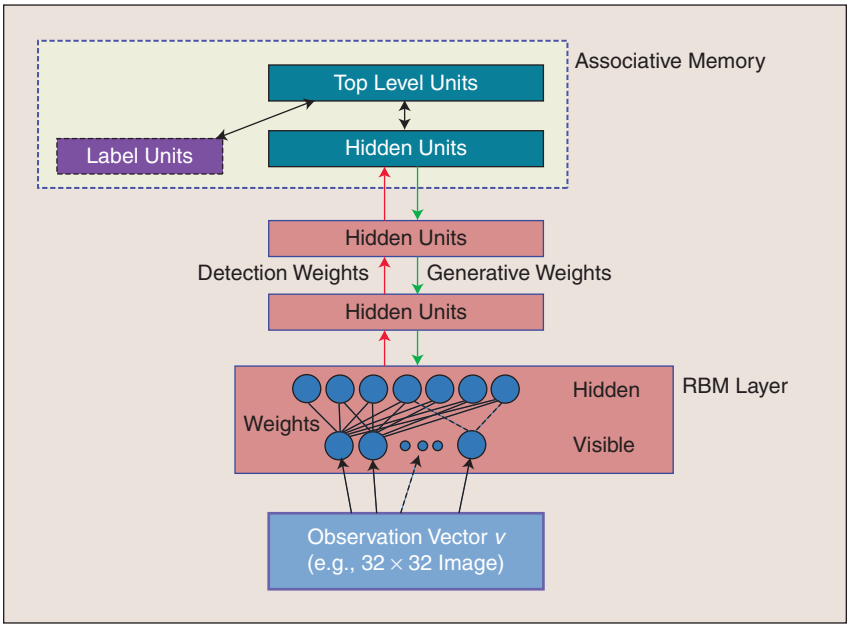


FIGURE 3 Illustration of the Deep Belief Network framework.

At the top two layers, the weights are tied together, such that the output of the lower layers provides a reference clue or link for the top layer to "associate" with its memory contents. We often encounter problems where discriminative performance is of ultimate concern, e.g. in classification tasks. A DBN may be fine tuned after pre-training for improved discriminative performance by utilizing labeled data through back-propagation. At this point, a set of labels is attached to the top layer (expanding the associative memory) to clarify category boundaries in the network through which a new set of bottom-up, recognition weights are learned. It has been shown in [16] that such networks often perform better than those trained exclusively with back-propagation. This may be intuitively explained by the fact that back-propagation for DBNs is only required to perform a local search on the weight (parameter) space, speeding training and convergence time in relation to traditional feed-forward neural networks.

Performance results obtained when applying DBNs to the MNIST handwritten character recognition task have demonstrated significant improvement over feedforward networks. Shortly after DBNs were introduced, a more thorough analysis presented in [17] solidified their use with unsupervised tasks as well as continuous valued inputs. Further tests in [18] [19] illustrated the resilience of DBNs (as well as other deep architectures) on problems with increasing variation.

The flexibility of DBNs was recently expanded [20] by introducing the notion of Convolutional Deep Belief Networks (CDBNs). DBNs do not inherently embed information about the 2D structure of an input image, i.e. inputs are simply vectorized formats of an image matrix. In contrast, CDBNs utilize the spatial relationship of neighboring pixels with the introduction of what are termed convolutional RBMs to provide a translation invariant generative model that scales well with high dimensional images. DBNs do not currently explicitly address learning the temporal relationships between observables, though there has been recent work in stacking temporal RBMs [22] or generalizations of these, dubbed temporal convolution machines [23], for learning sequences. The application of such sequence learners to audio signal processing problems, whereby DBNs have made recent headway [24], offers an avenue for exciting future research.

Static image testing for DBNs and CNNs occurs most commonly with the MNIST database [27] of handwritten digits and Caltech-101 database [28] of various objects (belonging to 101 categories). Classification error rates for each of the architectures can be found in [19] [20] [21]. A comprehensive and up-to-date performance comparison for various machine learning techniques applied to the MNIST database is provided in [27].

Recent works pertaining to DBNs include the use of stacked auto-encoders in place of RBMs in traditional DBNs [17] [18] [21]. This effort produced deep multi-layer neural network architectures that can be trained with the same principles as DBNs but are less strict in the parameterization of the layers. Unlike DBNs, auto-encoders use discriminative models from which the input sample space cannot be sampled by the architecture, making it more difficult to interpret what the network is capturing in its internal representation. However, it has been shown [21] that denoising auto-encoders, which utilize stochastic corruption during training, can be stacked to yield generalization performance that is comparable to (and in some cases better that) traditional DBNs. The training procedure for a single denoising autoencoder corresponds to the goals used for generative models such as RBMs.

IV. Recently Proposed Deep Learning Architectures

There are several computational architectures that attempt to model the neocortex. These models have been inspired by sources such as [42], which attempt to map various computational phases in image understanding to areas in the cortex. Over time these models have been refined; however, the central concept of visual processing over a hierarchical structure has remained. These models invoke the simple-to-complex cell organization of Hubel and Weisel [44], which was based on studies of the visual cortical cells of cats.

Similar organizations are utilized by CNNs as well as other deep-layered models (such as the Neocognitron [40] [41] [43] and HMAX [32] [45]), yet more "explicit" cortical models seek a stronger mapping of their architecture to biologically-inspired models. In particular, they attempt to solve problems of learning and invariance through diverse mechanisms such as temporal analysis, in which time is considered an inseparable element of the learning process.

One prominent example is Hierarchical Temporal Memory (HTM) developed at the Numenta Corporation [30] [33]. HTMs have a hierarchical structure based on concepts described in [39] and bear similarities to other work pertaining to the modeling of cortical circuits. With a specific focus on visual information representation, in an HTM the lowest level of the hierarchy receives its inputs from a small region of an input image. Higher levels of the hierarchy correspond to larger regions (or receptive fields) as they incorporate the representation constructs of multiple lower receptive fields. In addition to the scaling change across layers of the hierarchy, there is an important temporal-based aspect to each layer, which is created by translation or scanning of the input image itself.

During the learning phase, the first layer compiles the most common input patterns and assigns indices to them. Temporal relationships are modeled as probability transitions from one input sequence to another and are clustered together using graph partitioning techniques. When this stage of learning concludes, the subsequent (second) layer concatenates the indices of the current observed inputs from its children modules and learns the most common concatenations as an alphabet (another group of common input sequences, but at a higher level). The higher layer's characterization can then be provided as feedback down to the lower level modules. The lower level, in turn, incorporates this broader representation information into its own inference formulation. This process is repeated at each layer of the hierarchy. After a network is trained, image recognition is performed using the Bayesian belief propagation algorithm [46] to identify the most likely input pattern given the beliefs at the highest layer of the hierarchy (which corresponds to the broadest image scope). Other architectures proposed in the literature, which resemble HTMs, include the Hierarchical Quilted SOMs of Miller & Lommel [47] that employ two-stage spatial clustering and temporal clustering using self-organizing maps, and the Neural Abstraction Pyramid of Behnke [48].

A framework recently introduced by the authors for achieving robust information representation is the Deep SpatioTemporal Inference Network (DeSTIN) model [26]. In this framework, a common cortical circuit (or node) populates the entire hierarchy, and each of these nodes operates independently and in parallel to all other nodes.

TABLE I Summary of mainstream deep machine learning approaches.

APPROACH (ABBREVIATION)	UNSUPERVISED PRE-TRAINING?	GENERATIVE VS. DISCRIMINATIVE	NOTES
CONVOLUTIONAL NEURAL NETWORKS (CNNs)	NO	DISCRIMINATIVE	UTILIZES SPATIAL/TEMPORAL RELATIONSHIPS TO REDUCE LEARNING REQUIREMENTS
DEEP BELIEF NETWORKS (DBNs)	HELPFUL	GENERATIVE	MULTI-LAYERED RECURRENT NEURAL NETWORK TRAINED WITH ENERGY MINIMIZING METHODS
STACKED (DENOISING) AUTO-ENCODERS	HELPFUL	DISCRIMINATIVE (DENOISING ENCODER MAPS TO GENERATIVE MODEL)	STACKED NEURAL NETWORKS THAT LEARN COMPRESSED ENCODINGS THROUGH RECONSTRUCTION ERROR
HIERARCHICAL TEMPORAL MEMORY	NO	GENERATIVE	HIERARCHY OF ALTERNATING SPATIAL RECOGNITION AND TEMPORAL INFERENCE LAYERS WITH SUPERVISED LEARNING METHOD AT TOP LAYER
DEEP SPATIOTEMPORAL INFERENCE NETWORK (DESTIN)	NO	DISCRIMINATIVE	HIERARCHY OF UNSUPERVISED SPATIAL-TEMPORAL CLUSTERING UNITS WITH BAYESIAN STATE-TO-STATE TRANSITIONS AND TOP-DOWN FEEDBACK

This solution is not constrained to a layer-by-layer training procedure, making it highly attractive for implementation on parallel processing platforms. Nodes independently characterize patterns through the use of a belief state construct, which is incrementally updated as the hierarchy is presented with data.

This rule is comprised of two constructs: one representing how likely system states are for segments of the observation, $P(observation \mid state)$, and another representing how likely state to state transitions are given feedback from above, $P(subsequent\ state \mid state, feedback)$. The first construct is unsupervised and driven purely by observations, while the second, modulating the first, embeds the dynamics in the pattern observations. Incremental clustering is carefully applied to estimate the observation distribution, while state transitions are estimated based on frequency. It is argued that the value of the scheme lies in its simplicity and repetitive structure, facilitating multi-modal representations and straightforward training.

Table 1 provides a brief comparison summary of the mainstream deep machine learning approaches described in this paper.

V. Deep Learning Applications

There have been several studies demonstrating the effectiveness of deep learning methods in a variety of application domains. In addition to the MNIST handwriting challenge [27], there are applications in face detection [10] [51], speech recognition and detection [12], general object recognition [9], natural language processing [24], and robotics. The reality of data proliferation and abundance of multimodal sensory information is admittedly a challenge and a recurring theme in many military as well as civilian applications, such as sophisticated surveillance systems. Consequently, interest in deep machine learning has not been limited to academic research. Recently, the Defense Advanced Research Projects Agency (DARPA) has announced a research program exclusively focused on deep learning [29]. Several private organizations, including Numenta [30] and Binatix [31], have focused their attention on commercializing deep learning technologies with applications to broad domains.

VI. The Road Ahead

Deep machine learning is an active area of research. There remains a great deal of work to be done in improving the learning process, where current focus is on lending fertile ideas from other areas of machine learning, specifically in the context of dimensionality reduction. One example includes recent work on sparse coding [57] where the inherent high dimensionality of data is reduced through the use of compressed sensing theory, allowing accurate representation of signals with very small numbers of basis vectors. Another example is semi-supervised manifold learning [58] where the dimensionality of data is reduced by measuring the similarity between training data samples, then projecting these similarity measurements to lower-dimensional spaces. In addition, further inspiration and techniques may be found from evolutionary programming approaches [59, 60] where conceptually adaptive learning and core architectural changes can be learned with minimal engineering efforts.

Some of the core questions that necessitate immediate attention include: how well does a particular scheme scale with respect to the dimensionality of the input (which in images can be in the millions)? What is an efficient framework for capturing both short and long-term temporal dependencies? How can multimodal sensory information be most naturally fused within a given architectural framework? What are the correct attention mechanisms that can be used to augment a given deep learning technology so as to improve robustness and invariance to distorted or missing data? How well do the various solutions map to parallel processing platforms that facilitate processing speedup?

While deep learning has been successfully applied to challenging pattern inference tasks, the goal of the field is far beyond task-specific applications. This scope may make the comparison of various methodologies increasingly complex and will likely necessitate a collaborative effort by the research community to address. It should also be noted that, despite the great prospect offered by deep learning technologies, some domain-specific tasks may not be directly improved by such schemes. An example

is identifying and reading the routing numbers at the bottom of bank checks. Though these digits are human readable, they are comprised of restricted character sets which specialized readers can recognize flawlessly at very high data rates [49]. Similarly, iris recognition is not a task that humans generally perform; indeed, without training, one iris looks very similar to another to the untrained eye, yet engineered systems can produce matches between candidate iris images and an image database with high precision and accuracy to serve as a unique identifier [50]. Finally, recent developments in facial recognition [51] show equivalent performance relative to humans in their ability to match query images against large numbers of candidates, potentially matching far more than most humans can recall [52]. Nevertheless, these remain highly specific cases and are the result of lengthy feature engineering optimization processes (as well as years of research) that do not map to other, more general applications. Furthermore, deep learning platforms can also benefit from engineered features while learning more complex representations which engineered systems typically lack.

Despite the myriad of open research issues and the fact that the field is still in its infancy, it is abundantly clear that advancements made with respect to developing deep machine learning systems will undoubtedly shape the future of machine learning and artificial intelligence systems in general.

References

[1] R. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press, 1957.

[2] R. Duda, P. Hart, and D. Stork, *Pattern Recognition*, 2nd ed. New York: Wiley-Interscience, 2000.

[3] T. Lee and D. Mumford, "Hierarchical Bayesian inference in the visual cortex," *J. Opt. Soc. Amer.*, vol. 20, pt. 7, pp. 1434–1448, 2003.

[4] T. Lee, D. Mumford, R. Romero, and V. Lamme, "The role of the primary visual cortex in higher level vision," *Vision Res.*, vol. 38, pp. 2429–2454, 1998.

[5] G. Wallis and H. Bülthoff, "Learning to recognize objects," *Trends Cogn. Sci.*, vol. 3, no. 1, pp. 23–31, 1999.

[6] G. Wallis and E. Rolls, "Invariant face and object recognition in the visual system," *Prog. Neurobiol.*, vol. 51, pp. 167–194, 1997.

[7] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[9] F.-J. Huang and Y. LeCun, "Large-scale learning with SVM and convolutional nets for generic object categorization," in *Proc. Computer Vision and Pattern Recognition Conf. (CVPR'06)*, 2006.

[10] B. Kwolek, "Face detection using convolutional neural networks and Gabor filters," in *Lecture Notes in Computer Science*, vol. 3696. 2005, p. 551.

[11] F. H. C. Tivive and A. Bouzerdoum, "A new class of convolutional neural networks (SICoNNets) and their application of face detection," in *Proc. Int. Joint Conf. Neural Networks*, 2003, vol. 3, pp. 2157–2162.

[12] S. Sukittanon, A. C. Surendran, J. C. Platt, and C. J. C. Burges, "Convolutional networks for speech detection," *Interspeech*, pp. 1077–1080, 2004.

[13] Y.-N. Chen, C.-C. Han, C.-T. Wang, B.-S. Jeng, and K.-C. Fan, "The application of a convolution neural network on face and license plate detection," in *Proc. 18th Int. Conf. Pattern Recognition (ICPR'06)*, 2006, pp. 552–555.

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

[15] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, pp. 1771–1800, 2002.

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

[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)*. 2007, pp. 153–160.

[18] M. Ranzato, F. J. Huang, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. Computer Vision and Pattern Recognition Conf.*, 2007.

[19] 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 *Proc. 24th Int. Conf. Machine Learning (ICML'07)*, 2007, pp. 473–480.

[20] H. Lee, R. Grosse, R. Ranganath, and A. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. 26th Int. Conf. Machine Learning*, 2009, pp. 609–616.

[21] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Machine Learning (ICML'08)*, 2008, pp. 1096–1103.

[22] I. Sutskever and G. Hinton, "Learning multilevel distributed representations for high-dimensional sequences," in *Proc. 11th Int. Conf. Artificial Intelligence and Statistics*, 2007.

[23] A. Lockett and R. Miikkulainen, "Temporal convolution machines for sequence learning," Dept. Comput. Sci., Univ. Texas, Austin, Tech. Rep. AI-09-04, 2009.

[24] H. Lee, Y. Largman, P. Pham, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in Neural Information Processing Systems 22 (NIPS'09)*, 2009.

[25] H. Mobahi, R. Collobert, and J. Weston, "Deep learning from temporal coherence in video," in *Proc. 26th Annu. Int. Conf. Machine Learning*, 2009, pp. 737–744.

[26] I. Arel, D. Rose, and B. Coop, "DeSTIN: A deep learning architecture with application to high-dimensional robust pattern recognition," in *Proc. 2008 AAAI Workshop Biologically Inspired Cognitive Architectures (BICA)*.

[27] The MNIST database of handwritten digits [Online]. Available: http://yann.lecun.com/exdb/mnist/

[28] Caltech 101 dataset [Online]. Available: http://www.vision.caltech.edu/Image_Datasets/Caltech101/

[29] http://www.darpa.mil/IPTO/solicit/baa/BAA09-40_PIP.pdf

[30] http://www.numenta.com

[31] http://www.binatix.com

[32] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 3, pp. 411–426, 2007.

[33] D. George, "How the brain might work: A hierarchical and temporal model for learning and recognition," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2008.

[34] T. Dean, G. Carroll, and R. Washington, "On the prospects for building a working model of the visual cortex," in *Proc. Nat. Conf. Artificial Intelligence*, 2007, vol. 22, p. 1597.

[35] T. Dean, "A computational model of the cerebral cortex," in *Proc. Nat. Conf. Artificial Intelligence*, 2005, vol. 20, pp. 938–943.

[36] T. S. Lee and D. Mumford, "Hierarchical Bayesian inference in the visual cortex," *J. Opt. Soc. Amer. A*, vol. 20, no. 7, pp. 1434–1448, 2003.

[37] M. Szarvas, U. Sakai, and J. Ogata, "Real-time pedestrian detection using LIDAR and convolutional neural networks," in *Proc. 2006 IEEE Intelligent Vehicles Symp.*, pp. 213–218.

[38] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. 7th Int. Conf. Document Analysis and Recognition*, 2003, pp. 958–963.

[39] J. Hawkins and S. Blakeslee, *On Intelligence. Times Books*, Oct. 2004.

[40] K. Fukushima, "Neocognitron for handwritten digit recognition," *Neurocomputing*, vol. 51, pp. 161–180, 2003.

[41] K. Fukushima, "Restoring partly occluded patterns: A neural network model," *Neural Netw.*, vol. 18, no. 1, pp. 33–43, 2005.

[42] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, 1983.

[43] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, 1980.

[44] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, pp. 106–154, 1962.

[45] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nat. Neurosci.*, vol. 2, no. 11, pp. 1019–1025, 1999.

[46] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988.

[47] J. W. Miller and P. H. Lommel, "Biometric sensory abstraction using hierarchical quilted self-organizing maps," *Proc. SPIE*, vol. 6384, 2006.

[48] S. Behnke, *Hierarchical Neural Networks for Image Interpretation*. New York: Springer-Verlag, 2003.

[49] S. V. Rice, F. R. Jenkins, and T. A. Nartker, "The fifth annual test of OCR accuracy," *Information Sciences Res. Inst.*, Las Vegas, NV, TR-96-01, 1996.

[50] E. M. Newton and P. J. Phillips, "Meta-analysis of third-party evaluations of iris recognition," *IEEE Trans. Syst., Man, Cybern. A*, vol. 39, no. 1, pp. 4–11, 2009.

[51] M. Osadchy, Y. LeCun, and M. Miller, "Synergistic face detection and pose estimation with energy-based models," *J. Mach. Learn. Res.*, vol. 8, pp. 1197–1215, May 2007.

[52] A. Adler and M. Schuckers, "Comparing human and automatic face recognition performance," *IEEE Trans. Syst., Man, Cybern. B*, vol. 37, no. 5, pp. 1248–1255, 2007.

[53] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 328–339, 1989.

[54] K. Lang, A. Waibel, and G. Hinton, "A time-delay neural-network architecture for isolated word recognition," *Neural Netw.*, vol. 3, no. 1, pp. 23–44, 1990.

[55] R. Hadseel, 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*, 2008, pp. 628–633.

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

[57] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun, "Learning invariant features through topographic filter maps," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2009.

[58] J. Weston, F. Ratle, and R. Collobert, "Deep learning via semi-supervised embedding," in *Proc. 25th Int. Conf. Machine Learning*, 2008, pp. 1168–1175.

[59] K. A. DeJong, "Evolving intelligent agents: A 50 year quest," *IEEE Comput. Intell. Mag.*, vol. 3, no. 1, pp. 12–17, 2008.

[60] X. Yao and M. Islam, "Evolving artificial neural network ensembles," *IEEE Comput. Intell. Mag.*, vol. 2, no. 1, pp. 31–42, 2008.