Biologically Inspired Cognitive Architectures II:
Papers from the AAAI Fall Symposium (FS-09-01)

DeSTIN: A Scalable Deep Learning Architecture with Application to High-Dimensional Robust Pattern Recognition

Itamar Arel and **Derek Rose** and **Robert Coop**
Department of Electrical Engineering and Computer Science
Machine Intelligence Lab
The University of Tennessee
1508 Middle Drive
Knoxville, TN 37996-2100

Abstract

The topic of deep learning systems has received significant attention during the past few years, particularly as a biologically-inspired approach to processing high-dimensional signals. The latter often involve spatiotemporal information that may span large scales, rendering its representation in the general case highly challenging. Deep learning networks attempt to overcome this challenge by means of a hierarchical architecture that is comprised of common circuits with similar (and often cortically influenced) functionality. The goal of such systems is to represent sensory observations in a manner that will later facilitate robust pattern classification, mimicking a key attribute of the mammal brain. This stands in contrast with the mainstream approach of pre-processing the data so as to reduce its dimensionality - a paradigm that often results in sub-optimal performance. This paper presents a Deep SpatioTemporal Inference Network (DeSTIN) - a scalable deep learning architecture that relies on a combination of unsupervised learning and Bayesian inference. Dynamic pattern learning forms an inherent way of capturing complex spatiotemporal dependencies. Simulation results demonstrate the core capabilities of the proposed framework, particularly in the context of high-dimensional signal classification.

Introduction

Recent neuroscience research findings suggest that the neocortex may comprise of a large number of identical building blocks (i.e. cortical circuits) that populate a hierarchical structure (A.J. Rockel and Powell 1980; Felleman and Essen 1991; Phillips and Singer 1997; Douglas and Martin 2004). Such structure allows mammals to effectively learn to represent sensory information, particularly in the context of capturing spatiotemporal dependencies. The core assumption is that by partitioning high-dimensional sensory signals into smaller segments and modeling those based on regularities in the observations, a scalable system emerges, which is capable of dealing with the virtually infinite (though structurally bound) amount of information mammals are exposed to over time. This paradigm inspired researchers in recent years to develop several approaches for such deep learning systems. Most notable are deep belief networks (DBNs) (G. E. Hinton and Teh 2006) and convolutional neural networks (Y. LeCun and et al. 1990). DBNs are probabilistic generative models that are composed of multiple layers of stochastic, latent variables; traditionally DBNs lack the ability to combine unsupervised learning with supervised learning in a manner that allows unlabeled observations to be learned and represented independently of labeled ones. Recent work by (H. Lee and Ng 2009) has made great strides in scaling both unsupervised and semi-supervised learning in DBNs, though training of these models remains computationally costly. Convolutional neural networks are discriminative connectionist models designed to operate directly on observed images without preprocessing. They have been proven robust in the presence of noise and (reasonable levels of) geometric distortion or transformation in the context of image classification. However, to the best of our knowledge, these existing deep learning schemes do not inherently combine spatial *and* temporal information in a scalable manner. Work on connectionist architectures which handle spatiotemporal relationships has been explored, though recognition results and scalability questions haven't been addressed (Chappelier and Grumbach 1998).

Physiologically supported by research in the visual area of the cortex (Lee and Mumford 2003), George and Hawkins (2005) have introduced a distinct generative Bayesian inference model that incorporates learning spatiotemporal dependencies. Object recognition results have shown the viability of their architecture, while Pearl's message passing interface for Bayesian networks provides scalability in specially constructed graphs (Pearl 1988). Work presented here has similar goals and formulation, though we propose that the generative model is unnecessary for utilizing sequence inferences and believe such inferences can be well employed in current control problems, particularly in a partially observable domain.

Processing high-dimensional signals has been a core challenge in many science and engineering applications over the last few decades. The main difficulty that arises with the high-dimensionality of signals, particularly in classification applications, is that learning complexity grows exponentially with linear increase in the signal dimensionality. This phenomenon is known as the *curse of dimensionality*. Thus, the mainstream approach for dealing with high-dimensional

inputs has been to pre-process the information in a manner that would reduce the dimensionality into that which can be effectively applied to a classification engine. This process is often referred to as feature extraction. As a result, it can be argued that the intelligence behind such systems shifts to the human-engineered feature extraction process, which at times can be challenging and highly application-dependent.

An alternative stand, which is more biologically-inspired, argues that the neocortex does not explicitly pre-process sensory signals, but rather allows them to propagate through a complex hierarchy of modules that incrementally learn to represent observations based on the regularities they exhibit (Barlow 1989). Under this assumption, it would be possible to train such a hierarchical network on a large set of observations and later extract signals from this hierarchical network to a simple classification engine for the purpose of robust pattern recognition. Robustness here refers to the ability to exhibit invariance to a diverse range of transformations and distortions, including noise, scale, rotation, and lighting conditions.

This paper presents a novel discriminative deep learning architecture that combines concepts from unsupervised learning for dynamic pattern representation together with Bayesian inference. This architecture, which we deem a Deep SpatioTemporal Inference Network, yields a highly scalable modeling system which is capable of effectively dealing with high-dimensional signals. Moreover, the spatiotemporal dependencies that exist within the observations are modeled inherently in an unguided manner. Each node (in a believed cortical circuit manner) models patterns (or sequences) it observes by means of clustering, while it constructs a belief state over this distribution of sequences using Bayesian inference. We demonstrate that information from the top layer of this hierarchical system can be extracted and utilized for the purpose of pattern classification. Simulation results indicate that the framework is highly promising, paving the way for various real-world machine learning applications.

Deep SpatioTemporal Inference Network (DeSTIN)

Our primary conjecture supporting the approach taken to a deep learning architecture with common circuit functionality is stated as follows.

Conjecture 1. *The sequential structure of observed data can be represented hierarchically, whereby long-term dependencies (between two events remote from each other in time) will not depend on precise temporal intervals (i.e. on the accurate timing of these events).*

In essence, this assumption translates to a very simple and general *a priori* on the structure of the data. Consequently, instead of a single homogeneous state abstraction variable, several "layers" of state abstractions, each working at a different time scale, are employed. The interpretation here is that long-term dependencies are insensitive to small timing variations, such that they can be represented using a coarse temporal scale. This scheme allows context information and credit information to be respectively propagated forward and backward more easily, lending the architecture to pipelined implementations. Multiple time scales are thus represented at every instant revealing regularities in the observations.

Probabilistic Foundation

We thus begin by deriving the basic belief update rule, which is identical for every node in the architecture. The belief of each node represents the distribution of likelihood over the sequences that it learns to represent. In other words, each node is allocated a predefined number of dynamic patterns, or sequences. We seek to derive an update rule that would map the current observation (o), belief state (b), and belief state of a higher layer node (c), to a new (updated) belief state (b'), such that

$$b'\left(s'\right) = \Pr\left(s'|o,b,c\right) = \frac{\Pr\left(s' \cap o \cap b \cap c\right)}{\Pr\left(o \cap b \cap c\right)}, \quad (1)$$

alternatively expressed as

$$b'\left(s'\right) = \frac{\Pr(o|s',b,c)\Pr\left(s'|b,c\right)\Pr\left(b,c\right)}{\Pr\left(o|b,c\right)\Pr\left(b,c\right)}. \quad (2)$$

Under the assumption that observations depend only on true state, or $\Pr(o|s',b,c) = \Pr(o|s')$, we can further simplify the expression such that

$$b'\left(s'\right) = \frac{\Pr(o|s')\Pr\left(s'|b,c\right)}{\Pr\left(o|b,c\right)}, \quad (3)$$

where $\Pr\left(s'|b,c\right) = \sum_{s\in S} \Pr\left(s'|s,c\right) b\left(s\right)$, yielding the belief update rule

$$b'\left(s'\right) = \frac{\Pr\left(o|s'\right) \sum_{s\in S} \Pr\left(s'|s,c\right) b\left(s\right)}{\sum_{s''\in S} \Pr\left(o|s''\right) \sum_{s\in S} \Pr\left(s''|s,c\right) b\left(s\right)}, \quad (4)$$

where S denotes the sequence set (i.e. belief dimension) such that the denominator term is a normalization factor. One interpretation of (4) would be that the static pattern similarity metric, $\Pr\left(o|s'\right)$, is modulated by a construct that reflects the system dynamics, $\Pr\left(s'|s,c\right)$. As such, the belief state inherently captures both spatial an temporal information. In our implementation, the belief state of the parent node, c, is chosen using the selection rule

$$c = \arg\max_s b_p(s) \quad (5)$$

where b_p is the belief distribution of the parent node. A closer look at eq. (4) reveals that there are two core constructs to be learned, $\Pr(o|s')$ and $\Pr(s'|s,c)$. We show that the former can be learned via online clustering while the latter is learned based on experience by adjusting of the parameters with each transition from s to s' given c. The result is a robust framework that autonomously (i.e. with no human engineered pre-processing of any type) learns to represent complex data patterns, such as those found in real-life robotics applications.

Observation Clustering

To form $\Pr(o|s')$, we utilized an online clustering algorithm with a finite set of centroids that represent each possible "state" (i.e. sequence or pattern progression) s for each node. The size of the centroid set is a defined parameter which is preferably larger than necessary to give each node sufficient representational capacity to model the observation primitives presented. We use a winner take all (WTA) approach to unsupervised clustering with Euclidean distance metric. Tests with the cosine similarity metric have been somewhat promising, though our results have been best with Euclidean distance. In higher dimensional spaces, we have found that online clustering with a WTA method can create idle clusters and improved the algorithm with an additional credit assignment factor.

Idle clusters form as a result of poor initial placement and observation frequency, where a cluster is "starved" if it is initially further from two observation clustering locations than a neighboring centroid. In this case, the closer centroid tends to pull to an intermediate location which remains closer than the starved centroid. To combat such a problem, idle centroids accumulate credit for every observation seen and, when selected, lose their credit. For every centroid x and observation o, distances are calculated as follows:

$$d_x = dist(x, o) = \|x - o\| \, starvation(x) \tag{6}$$

WTA centroid selection then simply performs $\arg\min_x d_x$ and repositions the centroid. The adjustment magnitude (i.e. step size) is augmented as a result of the prior update to the centroid, in a similar fashion to momentum. The intuition here is that larger movements of the centroid should preclude progressively smaller fine tuning or major shifts in potential observations, thus compelling a larger update on the next iteration. If we define the prior centroid update for centroid x to be Δ_x, the new step size, α_x may be calculated as

$$\rho = \frac{(x - o)\,(1 - \Delta_x)}{\|x - o\| \, \|\Delta_x\|} \tag{7}$$

$$\alpha_x = \exp\left((\beta - 1)\,\rho\right) \tag{8}$$

where $\beta > 1$. Further, we create a safeguard against interrupted data streams or other observation irregularities (temporary noise or jitter) by smoothing the step size over each centroid's relative time. Let α_x^t be the previous step size for centroid x. If the centroid is chosen by the WTA algorithm, we may update it's step size proportional to a transfer constant γ

$$\alpha_x^{t+1} = \gamma \alpha_x^t + (1 - \gamma)\,\alpha_x \tag{9}$$

The motivation for such care in smooth clustering is indicated by the recursive reliance of belief updates shown above. Our architecture is especially sensitive to perturbations in centroid positioning. If observations of similar origin skip from centroid to centroid, convergence of the belief update and overall stability of the architecture is called into question. Though we have created stable systems, we still believe a more biologically inspired memory modeling technique for clustering observations presents an avenue for future work (i.e. in centroid trails).

For every observation and state s', we calculate $\Pr(o|s')$ in the belief update rule by taking the observation distance from the centroid labeled s' and dividing by the sum of all distances from state set S. This is normalized to preserve the unit sum of probabilities.

Architectural Strengths and Weaknesses

We have constructed the system presented here in a manner which builds upon practical implementation benefits of hierarchical systems in general, ensuring that the architecture is both parallelizable and can be pipelined. It is not necessary for any two nodes to share information within any stage of the pipeline, such that each could perform clustering and belief update rules independently. Furthermore, delays between pipeline stages rely only upon the maximum belief state computation of the previous stage (i.e. layer). We believe the common functionality and limited memory sharing of each node in our architecture will lend itself well to implementation on GPUs.

Unsupervised clustering methods utilized in the DeSTIN architecture provide a data-driven construction approach. Regularities in observations alone are all that is necessary to build a usable model, and in conjunction with the belief update rule presented above, provide a very simple theoretical foundation for a learning architecture. In comparison with existing deep inference architectures, the DeSTIN requires no initial pre-training. Though performance may be improved through random centroid assignment, this isn't a necessity for unsupervised centroid convergence. This stands in contrast to DBNs, which must train in a greedy layer-wise fashion on initialization (G. E. Hinton and Teh 2006).

Every layer (and further, node) within the architecture trains concurrently, though further research is necessary to compare the number of training examples necessary for convergence as well as computation time for such convergence between the DeSTIN, convolutional neural networks, and DBNs. Both DBNs and convolutional networks work well for 2D images, though their use in multiple modalities becomes somewhat of an art due to the non-linear operations performed on their inputs. The DeSTIN may operate on multiple modalities, provided distance metrics for such inputs are producible for clustering purposes.

Simulation Results

A basic high-dimensional pattern recognition task was considered as a means of testing and evaluating the basic concept introduced in this paper. The task involved classifying patterns into one of three distinct letters (the letters 'A', 'B' and 'C' were chosen). The input image comprised of 32×32 pixels, where each pixel was binary encoded as either black or white. Inputs were presented in sequences such that each sequence involved introduced a letter that was motioned across the view area in a 'Z' shaped pattern for 20 pixel steps. Letters were chosen at random for presentation. We feel the pattern of presentation for these simple observations is less important, however further research into optimal presentation patterns (which may source biological features

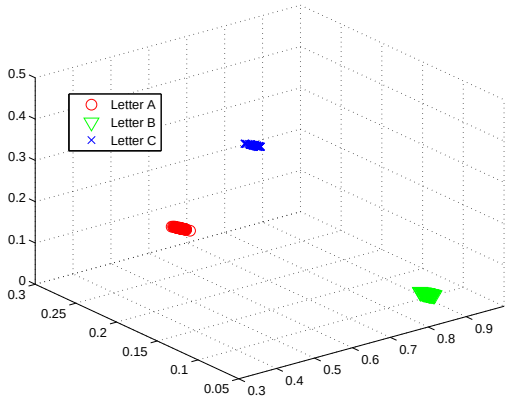

Figure 1: State space representation for the top layer node of the DeSTIN employed in the letters recognition testbench. Letters were presented with no noise

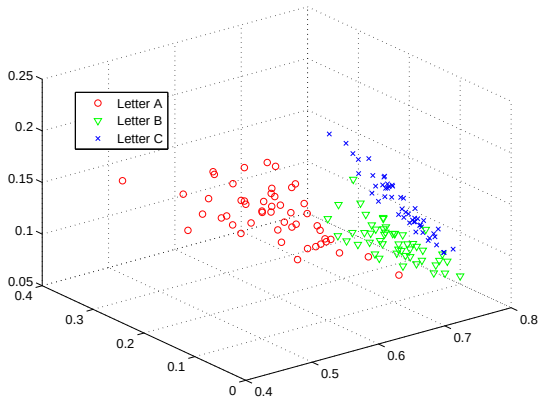

Figure 2: State space representation for the top layer node of the DeSTIN employed in the letters recognition testbench. Letters were presented with AWG noise (SNR 10dB)

such as saccading) could be necessary for learning more intricate objects.

The DeSTIN topology chosen consisted of 4 layers with the first layer hosting 64 nodes, each receiving a non-overlapping 4×4 patch of the image that is vectorized into a 16 element input. At each subsequent layer, there were one quarter of the nodes of the preceding layer, such that layer two hosted 16 nodes, layer three 4 nodes, and layer four hosted one node. The dimensionality of the belief states for layers one through four were 24, 12, 6 and 3, respectively. It is noted that as the layer index increases, the information compression rate increases, as reflected by a reduced dimensionality of the belief space.

Training the system was achieved by presenting the architecture with randomly selected letters and allowing the nodes to capture regularities in the observations in an unsupervised manner. In order to improve the overall process, during the first 300 presentations of letters the belief states were not updated at any node, allowing the clustering process at the first layer to begin converging prior to any adjusting of the belief constructs. This is not a mandatory apparatus, but rather a technique that accelerates the learning process as a whole.

Two cases were investigated. In the first, letters were presented with no noise or other forms of distortions applied (i.e. fixed patterns). In the second case, an identical topology was considered, whereby additive white Gaussian noise was applied to each image, with a signal to noise ratio (SNR) of 10 dB. This noise moves our observation space into the reals, highlighting a more general foundation for signal input. Figure 1 depicts the three-dimensional state space of the top layer node under no noise conditions. As can be observed, there are three clusters formed within this space that clearly correspond to the three letters presented at the lowest layer. Figure 2 illustrates the same state space for the case where random noise was added to each presented image (during both testing and training phases). The relatively high degree of noise resulted in a spreading of the three clusters, however linear separation was still attainable with close to 100% classification accuracy. These results show that the inference framework introduced, along with the specific methodology governing the learning process, are useful for unsupervised image classification. It is expected that these concepts can be directly applied to larger-scale problems in which the dimensionality of the images is higher and their information content much more complex.

Discussion

We have presented a novel deep inference architecture which draws upon the fundamentals of Bayes' theorem and unsupervised learning. Preliminary results are promising and clearly demonstrate an ability for the architecture to grow with high dimensional input in both a scalable and stable manner. Avenues for future work include applying this architecture to non-image based signals to accommodate other modalities, as well as recognized handwritten image databases (MNIST and Caltech 101). Moreover, we intend to investigate alternative distance measures that work well with high dimensional, real-world observations. As presented here, we currently use the top-layer belief states of the DeSTIN for classification purposes. Intuitively, this representation may not be appropriately fine grained for the classification problem at hand. As an example, consider identifying a cursive letter versus a printed one or differentiating a letter typed in different fonts. Extracting features from multiple layers through sampling may provide our architecture with a formalism that mimics salience detection. Evaluation of this output could point to new ways to present observations to the DeSTIN.

We argue that the inference framework and update rules introduced offer a robust scheme for capturing spatiotemporal dependencies. However, every component of that formulation can be realized using different technologies. For example, replacing the basic clustering method described

with a neural network or fuzzy logic based clustering may be found more appropriate for some applications. Representation of the system dynamics, which is achieved via the transition probability construct, can also be approximated rather than directly estimated, which should offer scalability proprieties as well as speed of convergence for large-scale problems.

In contrast with known deep inference architectures, such as deep belief networks, our scheme does not require any form of pre-training nor does it necessitate a layer-by-layer learning process, which would limit scalability and applicability to parallel computing platforms. Moreover, we consider the simplicity of the DeSTIN a primary strength for implementation, ease of extension, and scalability. The functions which are repeated at every node are succinct and easily implemented. Message passing is also straight forward, as no transformation functions are necessary - it is up to each node to contextualize information inherently. While we realize that the architecture presented here is perhaps overly simplified, we conjecture it practical and constructive.

References

A.J. Rockel, R. H., and Powell, T. 1980. The basic uniformity in structure of the neocortex. *Brain* 103:221–244.

Barlow, H. 1989. Unsupervised learning. *Neural Computation* 1:295–311.

Chappelier, J., and Grumbach, A. 1998. Rst: a connectionist architecture to deal with spatiotemporal relationships. *Neural Computation* 10:883–902.

Douglas, R., and Martin, K. 2004. Neuronal circuits of the neocortex. *Annual Review of Neuroscience* 27:419–451.

Felleman, D., and Essen, D. V. 1991. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* 1:1–47.

G. E. Hinton, S. O., and Teh, Y. 2006. A fast learning algorithm for deep belief nets. *Neural Computation* 18:1527–1554.

H. Lee, R. Grosse, R. R., and Ng, A. 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th International Conference on Machine Learning*.

Lee, T., and Mumford, D. 2003. Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America* 20:1434–1448.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Phillips, W., and Singer, W. 1997. In search of common foundations for cortical computation. *Behavioral and Brain Sciences* 20:657–722.

Y. LeCun, B. Boser, J. S. D., and et al. 1990. Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems* 2.