# How Can the Blind Men See the Elephant?

**Bonny Banerjee**
Institute for Intelligent Systems, and Dept. of Electrical & Computer Engineering
The University of Memphis, Memphis, TN 38152, USA
*bonnybanerjee@yahoo.com*

### Abstract

There is no denying the fact that AI's original aim of reproducing human-level intelligence has taken a back seat in favor of the development of practical and efficient systems for important but narrow domains under numerous umbrellas. While the importance of perception for intelligence is now well-understood, a major hurdle is to discover the appropriate representation and processes that can seamlessly support low-level perception and high-level cognition in a computational architecture. There is no shortage of cognitive architectures (see Samsonovich 2010 for a catalog), however, principled design is scarce. In this paper, we explicate our position and report on our ongoing investigations on representation and processes for developing an intelligent agent from first principles.

## The Goal

We posit that the world is composed of interactions between spatiotemporal objects in space and time at multiple levels of abstraction, where an object is defined as a set of transformation-invariant signals repeatedly coincident in space and time. The extent of this coincidence in space/time determines the spatial/temporal size of the object. The signals might emanate from sensors in any perceptual modality as well as from multiple modalities. Examples of spatiotemporal objects include visuospatial objects (e.g., chair, human), actions (e.g., running), events (e.g., party, exam, war), audio objects (e.g., phonemes, spoken words), and so on. Then, in order to live in this world, the primary goal of an intelligent agent is to be able to discover coherent spatiotemporal objects at the highest level of abstraction from sensory data in different modalities using all available resources. Examples of resources include, but not limited to, sensors, actuators, short- and long-term memory, sensory data, data labels, rewards and punishments.

As a concrete example, consider an agent responsible for the surveillance of a multi-story university building comprising of different kinds of spaces, such as, classrooms, labs, offices, restrooms, storerooms, corridors, elevators and staircases. In order for the agent to detect any abnormality in any space, it is imperative for it to know the norm in each of these spaces. To realize the difficulty of the problem, note that each classroom is different; in fact, the same classroom is different at different times of the day with different students sitting in different locations, different instructors teaching in different ways, and so on. How would the agent discover the invariance or norm across all these different scenarios, many of which it has never perceived before, with minimal supervision? That is the crux to discovering coherent spatiotemporal objects at the highest level of abstraction from sensory data. It is noteworthy that a typical human at a young age, possibly even higher animals, can recognize an environment very quickly; the tricks to attain such expertise are never taught explicitly in school or at home.



## Our Approach

The story of six blind men who went to "see" an elephant and ended up in complete disagreement as each touched a different part of its body, is well-known. They will be able to "see" the elephant if they can assimilate their perceptions of the different parts into a coherent whole. One way of achieving this is by crudely assuming the important functions of an elephant and explaining how each of their perceived parts contribute to performing those functions. The presumed functions then becomes a crucial factor, hence it is of utmost importance to select them judiciously. Our research on understanding how the brain works, keeping the goal stated in the last section in perspective, follows a similar vein, where different aspects of the brain circuitry is analogous to the different parts of the elephant.

A natural language captures in an unpremeditated fashion the many, if not all, thoughts that arise in the collective human minds intent on communicating in that language. Natural languages have evolved over a long time and are continuously evolving. All languages have lexical categories or parts of speech (e.g., noun, verb) with different variations that classifies all words in the language. These lexical categories are our presumed functions. We want to understand, in a principled way, how each of these lexical categories might be implemented in the brain networks such that they may be perceived from raw signals in different modalities. In particular, we are interested in five lexical categories in English – noun, verb, adjective, adverb and preposition. Being successful in this endeavor will enable us to build intelligent agents that can learn a large vocabulary of words as

percepts from raw signals where each percept represents the invariance for a class of spatiotemporal objects. If these percepts are labeled with words, such an agent will be able to understand the words in an embodied and grounded manner, and perform beyond symbol manipulation. In the rest of this paper, we present our framework, including representations and processes, guided by this approach, along with selected results and some intriguing issues.

## Framework

In our framework, an object $\mathcal{O}^{(\ell)}$, at any level of abstraction $\ell$, is composed of a finite set of invariant features $\{\mathcal{O}_1^{(k_1)}, \mathcal{O}_2^{(k_2)}, ...\mathcal{O}_n^{(k_n)}\}$ where each feature $\mathcal{O}_i^{(k_i)}$ is an object at a lower level, i.e. $k_i < \ell \; \forall i$, $i \neq j \nRightarrow k_i \neq k_j$. In order to design a minimalist model for discovering such objects, we speculate two hypotheses.

*Horizontal hypothesis*: The objectives of learning algorithms operating in the different perceptual cortices are similar. This common cortical algorithm hypothesis is supported by neuroscience evidence (Constantine-Paton and Law 1978; Metin and Frost 1989; von Melchner, Pallas, and Sur 2000; Bach-y-Rita and Kercel 2003).

*Vertical hypothesis*: The brain implements a small set of computations that are recursively executed at multiple stages of processing, from low-level perception (suited for tasks such as numeral recognition, speech recognition) to high-level cognition (suited for tasks such as medical diagnosis, criminal investigation). Besides being a useful design constraint, it helps identify canonical computations in the brain, existence of which has been indicated by research in sensory systems (Rust et al. 2005; Kouh and Poggio 2008; David et al. 2009; Douglas and Martin 2010).

### Architecture

Our model is implemented in a neural network architecture consisting of a hierarchy of layers of canonical computational units called nodes (see Fig. 1a) where each layer corresponds to a level of abstraction (Banerjee and Dutta 2013c). Each layer $\ell$ consists of two sublayers – simple neurons in the lower sublayer $S^{(\ell)}$ and complex neurons in the higher sublayer $C^{(\ell)}$ (see Fig. 1b).[1] Thus, our architecture is a cascade of alternating simple and complex sublayers, similar to a number of multilayered neural models, such as Neocognitron (Fukushima 2003), HMAX (Serre et al. 2007) and convolutional neural networks (LeCun and Bengio 1995), though they do not necessarily have a node structure. Neurons in a node are connected to those in the neighboring nodes in the same layer, one layer above and one layer below by lateral, bottom-up and top-down connections respectively. The lowest layer in the hierarchy receives input from external stimuli varying in space and time.

Each neuron has a spatial receptive field (RF) and a temporal RF, both of fixed sizes. The size of a stimulus it optimally responds to may be less than or equal to its spatial and temporal RF sizes. All neurons in a sublayer have same sized RFs. The size of spatial RF of a simple neuron in $S^{(\ell)}$ is defined by the number of nodes in its lower layer, $C^{(\ell-1)}$, reporting to it at any time instant. A neuron in layer $\ell$ samples the input stream every $\tau^{(\ell)}$ instants of time, where $\tau^{(\ell)}$ is referred to as the temporal RF size of the neuron. Complex neurons in $C^{(\ell)}$ sample the input at a lower frequency than simple neurons in $S^{(\ell)}$, i.e. $\tau^{(C^{(\ell)})} > \tau^{(S^{(\ell)})}$. Conceptually, a neuron in $C^{(\ell)}$ fails to distinguish the temporal sequence of events occurring within $\tau^{(C^{(\ell)})}$ instants of time, and hence considers all of those events to occur simultaneously (Dutta and Banerjee 2013). However, a neuron in $S^{(\ell)}$ can keep track of the temporal sequence of events occurring within $\tau^{(C^{(\ell)})}$ time instants due to its higher sampling frequency. The feedforward weights leading to a neuron encode a set. For a simple neuron, such a set comprises a feature while for a complex neuron, the set comprises a transformation. Optimal response of a simple neuron indicates the existence of the feature in the input while that of a complex neuron indicates the existence of the transformation in the input. Functionally, a node is a set of spatial/temporal filters, all of which are applied to the same location in the input data.

A multitude of functions have been attributed to lateral connections in the cerebral cortex. They run both within a particular cortical area and between different cortical areas (Gilbert and Wiesel 1983), and may be divergent or convergent (Rockland and Lund 1983). Lateral connections in the visual cortex interconnect columns of orthogonal orientation specificity (Matsubara, Cynader, and Swindale 1987) as well as similar orientation specificity (Gilbert and Wiesel 1989), and produce both excitation and inhibition (Hirsch and Gilbert 1991).

In our architecture, lateral connections among simple neurons within a node encode temporal transition probabilities (see Fig. 1c). A simple layer encodes temporal correlations of spatial changes; such strong correlations may be abstracted as causal knowledge. For example, start button in remote is pressed at time $t_1$ (spatial change encoded by a simple neuron), TV turned on at $t_2$ (spatial change encoded by another simple neuron). This correlation is directional in time ($t_1 < t_2$). The lateral connections among complex neurons across nodes encode spatial transition probabilities (see Fig. 1a). A complex layer encodes spatial correlations of temporal changes; for example, backrest of chair is moved at location $x_1$ (temporal change encoded by a complex neuron), seat of chair is moved at $x_2$ (temporal change encoded by another complex neuron). Their strong correlation may be abstracted as structural knowledge that the backrest and seat belong to the same object called "chair". This correlation is non-directional. Without lateral connections, each neuron will merely learn a set of features without their relative locations in space or time.

### Processes

The processes include algorithms for learning, from data, the neural architecture and inferring from it. We posit that the brain runs a relentless cycle of $Surprise \rightarrow Explain \rightarrow$

[1] The terms "simple" and "complex" neurons are used due to their functional resemblance to simple and complex cells in V1. Our architecture is not designed to model any part of the brain; however, similarities, in particular with neocortex, are inevitable.

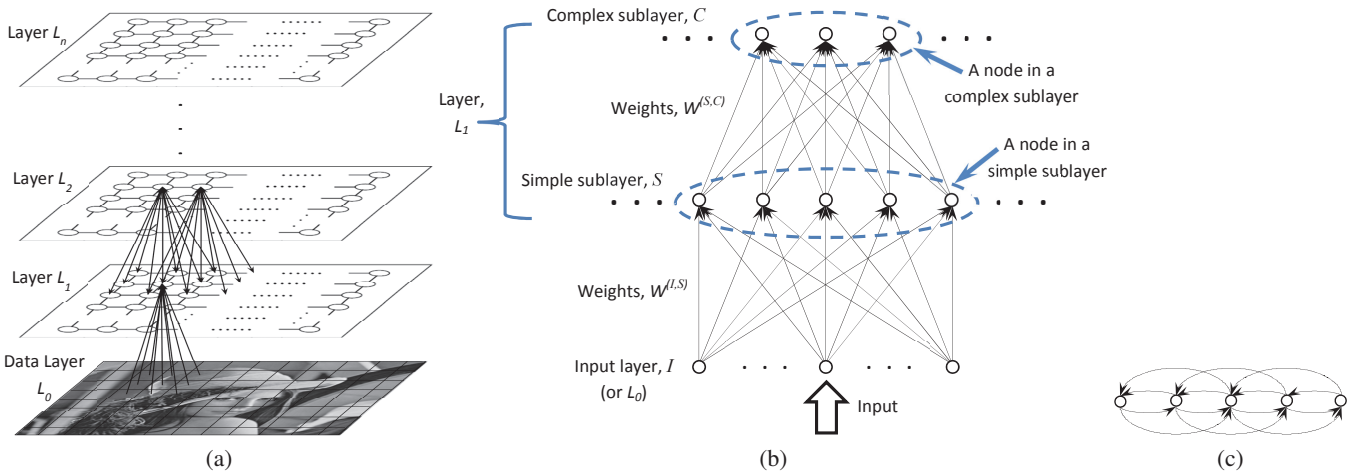



Figure 1: The neural network architecture used to implement our model. Reproduced from (Banerjee and Dutta 2013c). (a) Each layer $L_i$ is a pair of simple and complex sublayers. Circles denote nodes. (b) Feedforward connections from a simple to a complex sublayer node. Circles denote neurons. $W^{(I,S)}$ are learned to encode spatial sets or features in $S$. $W^{(S,C)}$ are learned to encode temporal sets or transformations in $C$. (c) Lateral connections in $S$ within a node. These lateral weights $W^{(S,S)}$ in conjunction with $W^{(S,C)}$ are modeled to learn sequences.

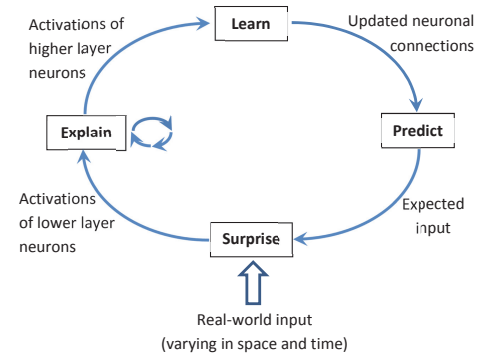
$Learn \rightarrow Predict$ involving the real external world and its internal model. Our perceptual organs sample the environment at regular intervals of time. To maintain a true correspondence between the real world and its internal model, the stimulus obtained at any sampled instant must be explained. By learning from explanation, the internal model is kept up-to-date with the nonstationary data distribution of the external world. An accurate internal model facilitates correct predictions (or expectations), thereby allowing us to act proactively and efficiently in the real world. When an expectation is violated, surprise occurs.

We use a general-purpose computational model, called SELP, for learning the neural architecture from space- and time-varying data in an unsupervised and online manner from surprises in the data (Banerjee and Dutta 2013c). Given streaming data, this model learns invariances using four functions – detect any unexpected or ***S***urprising event, ***E***xplain the surprising event, ***L***earn from its explanation, ***P***redict or expect the future events – and hence the name. In accordance with our vertical hypothesis, these four comprise the set of computations recursively carried out layer by layer. However, the relative amount of time spent on them might vary between layers (see Fig. 2). Typically, in lower layers, the explanation cycle runs within the prediction cycle at a faster time scale while it is the opposite in higher layers. Explanation in the lower perceptual levels can be crude as pixel-level accuracy in reconstructing a scene is seldom required. In contrast, higher cognitive levels have to explain and reason with complex scenarios which often require deliberation involving distal knowledge that cannot be executed quickly. On the other hand, accurate longer term predictions in the higher levels are seldom required. Since transformations in the lower levels are learned through predictions, hence predictive accuracy is crucial. Task requirements can always override these typical behaviors.

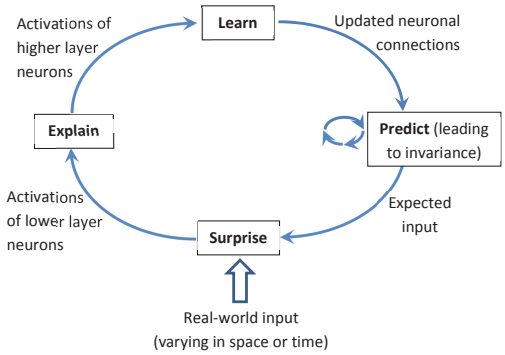
The prediction cycle in our model bears resemblance to predictive coding (Rao and Ballard 1999; Lee and Mumford 2003; Jehee et al. 2006; Bar 2007; Friston 2008; Spratling 2011; Chalasani and Principe 2013) which has been claimed to be employed in different parts of the brain, such as retina (Srinivasan, Laughlin, and Dubs 1982; Hosoya, Baccus, and Meister 2005), visual pathway (Fenske et al. 2006; Bar and others 2006; Hesselmann et al. 2010), auditory pathway (Vuust et al. 2009; Denham and Winkler 2006; Hesselmann et al. 2010; Winkler et al. 2012; Wacongne, Changeux, and Dehaene 2012), mirror neuron system (Kilner, Friston, and Frith 2007), and even in multisensory areas (van Wassenhove, Grant, and Poeppel 2005).

**Salient properties of SELP.** They are as follows:

*1. Learns efficiently.* Predictive coding is a very efficient paradigm for learning from streaming data as it learns selectively in time. It learns only when a prediction error or surprise occurs, unlike traditional learning algorithms that continue to learn from data all the time. However, predictive coding models, like traditional learning algorithms, learn from data in the entire space. Our model learns from spatiotemporal data selectively in time and space (Banerjee and Dutta 2013a). At any instant, it operates on input which is the change in the state of data between the current and last sampling instants as opposed to the entire state of data. This allows the model to concentrate on spatial regions of most significant changes and ignore those with minimal changes, thereby allowing all computational resources to be deployed for learning from a small but most interesting space of the data at any instant. Pursuit of efficiency leads to emergence of attention, inhibition of return, fixations and saccades in our model.




(a) SELP prediction cycle




(b) SELP explanation cycle

Figure 2: The SELP cycles. See text for details. Reproduced from (Banerjee and Dutta 2013c).

*2. Learns the relations in the input.* Our generative model does not generate the input but the relations in the input both in space and time. Higher level features are abstracted from correlations of lower layer neuronal activations (Dutta and Banerjee 2013). Such features are invariant to the absolute nature of the input. Since temporal (spatial) correlations of spatial (temporal) changes may be abstracted as causal (structural) knowledge, SELP is intrinsically designed to discover such knowledge. Learning from correlations leads to emergence of iconic and echoic memories and neural adaptation (Banerjee and Dutta 2013c).

*3. Learns shallow first, inferences deep first.* If higher layer features are abstracted from lower layer ones, the former will be unstable if the latter are. Thus, learning features layer-by-layer starting with the lowest or most shallow layer is widely followed in deep architectures. The higher features have a more global view of the world compared to their lower counterparts. Hence, inference from a higher level is expected to be more robust. In our model, the activations are propagated to the higher layers until a layer emerges as the winner. Since activations are due to surprises, the layer with the sparsest activations is considered the winner. This layer then dictates its lower layers to explain the input, if required.

*4. Learns discriminative features by explanation.* Features that discriminate between two objects or two classes of objects are learned in our model by first seeking them out through the process of explanation. Often discriminative features are salient, as in dog vs. man. However, when discriminative features are not so salient, as in identical twin siblings, a compute-intensive search in the space of features is required which is achieved by the SELP explanation cycle. This search may be implemented using orthogonal matching pursuit (OMP) (Pati, Rezaiifar, and Krishnaprasad 1993), an iterative algorithm widely used to compute the coefficients for the features in a generative model. OMP is a generalization of coefficient computation in spherical clustering as, for an input, the latter requires only one feedforward pass of OMP to determine the coefficient. Thus, the same algorithm can support a single feedforward pass to identify salient discriminative features as well as an iterative process to seek out the not-so-salient ones.

*5. Gets over the blind faith of supervised learning.* The implicit assumption in supervised learning is that data labels are generated from a source with no unreliability or noise. If the source is noisy, which is often the case in real world, the result of training with blind faith on the labels is worse than learning with no label. Our model treats labels as information in just another modality which may (or may not) provide consistent cues towards discovering an object.

Our algorithms for a simple layer are briefly described here. Details are in (Banerjee and Dutta 2013a; 2013b; 2013c). At each sampling instant, our model predicts the change in the data. If $X(t)$ is the state of the data at time $t$, the input to the model at $t$ is:

$$\Delta X(t) = X(t) - X(t-1) \tag{1}$$

The model's predicted input is $\Delta\hat{X}(t) = \hat{X}(t) - X(t-1)$. Hence, the predicted state of the data $\hat{X}(t)$ can be computed given the predicted input and the state of the data at the last time instant.

**Neuron.** Activation of $i^{th}$ simple neuron in $L_1$ at time $t$ is:

$$A_i^{(S)}(t) = \sum_j A_j^{(I)}(t) \times W_{ji}^{(I,S)}(t) \tag{2}$$

where $W_{ji}^{(I,S)}(t)$ is the weight or strength of connection from the $j^{th}$ neuron in input layer $I$ (or $L_0$) to the $i^{th}$ neuron in simple sublayer $S$ at $t$, and $A_j^{(I)}$ is the activation of $j^{th}$ neuron in $I$. Activation of a $S$ neuron denotes the strength of the feature, encoded by the feedforward weights leading to it, in the input. A simple neuron acts as a suspicious coincidence detector in space.

A complex neuron integrates activations from presynaptic simple neurons over its temporal RF and fires if the integrated input crosses its threshold. Activations of complex neurons in sublayer $C$ at time $t$ are:

$$A^{(C)}(t) = \sum_{h=t_0}^{t} A^{(S)}(h) \times W^{(S,C)}(h) \tag{3}$$

where $t_0$ is a time instant from when the complex neurons start integrating, $t - t_0 \leq \tau^{(C)}$. A complex neuron acts as a temporal coincidence detector. Each column in $W^{(I,S)}$ and $W^{(S,C)}$ is normalized to have unit norm. Operation in the complex layer is detailed in (Dutta and Banerjee 2013).

**Surprise.** A surprise is evoked when the actual input does not match the model's expected or predicted input. The input layer neurons are activated as a direct response to surprise which is computed as follows:

$$A^{(I)}(t) = \Delta X(t) - \Delta \hat{X}(t) \tag{4}$$

Note that this expression is the same as $X(t) - \hat{X}(t)$, i.e. the difference between the actual and predicted states of data. It has been conjectured that responding to a class of surprises is a basic function of individual neurons (Fiorillo 2008; Gill et al. 2008; Egner, Monti, and Summerfield 2010). Meyer and Olson (2011) have shown that inferotemporal neurons respond much more strongly to unpredicted transitions than to predicted ones. Gill et al. (2008) have shown that auditory neurons encode stimulus surprise as opposed to intensity or intensity changes.

**Explain.** At any sampling instant, the final activation $A^{(S)}$ of simple neurons is the sum of predicted activations $\hat{A}^{(S)}$ and the activations $\overline{A}^{(S)}$ required to explain the surprise or prediction error. In general, explanation may be construed as constructing a story or composite hypothesis that best accounts for the surprise. Here a simpler case is considered where explanation is construed as reconstruction of the surprise $A^{(I)}$ using the learned features and their activations $\overline{A}^{(S)}$, by minimizing the following loss function:

$$\mathcal{E}_{recon}(A^{(I)}|W^{(I,S)}) \equiv \frac{1}{2}\|A^{(I)} - W^{(I,S)} \times \overline{A}^{(S)}\|_2^2 \tag{5}$$

$$s.t.\ \ \|\overline{A}^{(S)}\|_0 \leq n$$

where $\|\overline{A}^{(S)}\|_0 \equiv \#\{i : \overline{A}_i^{(S)} \neq 0\}$, $n$ is a positive integer, and each column of $W^{(I,S)}$ is a feature that has been normalized to have unit norm. The condition on $\overline{A}^{(S)}$ constrains the maximum number of features used in the reconstruction, thereby inducing sparsity. Since $n$ is less than the number of available features, $\overline{A}^{(S)}$ may be computed using OMP.

Therefore, the final activation $A^{(S)}$ at time $t$ is:

$$A^{(S)}(t) = \begin{cases} \hat{A}^{(S)}(t), & \text{if there is no surprise} \\ & \text{i.e., } A^{(I)}(t) = 0 \\ \hat{A}^{(S)}(t) + \overline{A}^{(S)}(t), & \text{otherwise} \end{cases} \tag{6}$$

If there is no surprise, explanation is not required and the SELP cycle turns fast. If the surprise is large, the explanation might require longer time. Thus, the speed of execution of the SELP cycle is data dependent.

**Predict.** At any sampling instant, our model predicts the activation of the $i^{th}$ simple neuron for the next instant:

$$\hat{A}^{(S)}(t+1) = W^{(S,S)}(t) \times A^{(S)}(t) \tag{7}$$

where $W^{(S,S)}$ is the transition matrix encoded by the lateral weights in $S$. Then, the predicted change in input is:

$$\Delta \hat{X}(t+1) = W^{(I,S)}(t) \times \hat{A}^{(S)}(t+1) \tag{8}$$

The predicted input is: $\hat{X}(t+1) = X(t) + \Delta \hat{X}(t+1)$, though it is not required to be computed in our model.

**Learn.** There are two sets of weights, $W^{(I,S)}$ and $W^{(S,S)}$, to learn in a simple layer. The former is learned to minimize the reconstruction error while the latter to minimize prediction error. The prediction loss function is:

$$\mathcal{E}_{pred}(A^{(S)}|W^{(S,S)}) \equiv \frac{1}{2}\|A^{(S)} - \hat{A}^{(S)}\|_2^2 \tag{9}$$

where $\hat{A}^{(S)}$ is computed as in equation 7. Learning in the simple layer amounts to minimizing the explanation and prediction errors in conjunction, i.e. $\mathcal{E}_{recon} + \mathcal{E}_{pred}$ subject to $\|\overline{A}^{(S)}\|_0 \leq n$. There are different ways of computing this, see for example (Rao 1999; Chalasani and Principe 2013).

## Discussions

Our model was exposed to spatiotemporal data in different modalities. To illustrate how it learns from surprise, the following experiment was performed. Two images, that of Barbara and Lena, were presented to the input layer $I$ for t=1 through 49 and 50 through 100 iterations respectively. During the presentations, spatial structure in the images were learned by the lateral connections across nodes while the temporal pattern of presentation of the same image (t=1–49, 50–100) was learned by the lateral connections in each node. At any iteration, the current spatial and temporal structures learned in the network can be viewed from its expectations, as shown in Fig. 3. Learning of the new structure is noteworthy as image of Barbara was suddenly changed to that of Lena at t=50. The ghost of Barbara gradually disappears via a damped oscillatory behavior in the background of Lena. This phenomenon, akin to iconic memory, is due to the inability of lateral connections to forget the existing correlations immediately. The analogous phenomenon of echoic memory is observed when the same experiment is performed with audio stimuli.

To test feature learning in simple layer, as stimuli we used 17 videos recorded at different natural locations with a CCD camera mounted on a cat's head exploring its environment (Betsch et al. 2004). These videos provided a continuous stream of stimuli similar to what the cat's visual system is naturally exposed to, preserving its temporal structure. Using the objective in equation 5, the features learned in the feedforward connections by the simple layer is shown in Fig. 4. Qualitatively, the features belonged to three distinct classes of RFs – small unoriented features, localized and oriented Gabor-like filters, and elongated edge-detectors. Such features have been observed in macaque V1. Some transformations learned in the complex layer are shown in Fig. 5.

A generative model can be learned with weights constrained to be non-negative (see Fig. 4). We observe, learning with unconstrained weights often gives rise to complementary features, such as on-center-off-surround and off-center-on-surround, which raises the question – do we need to perceive instances and their complements in order to learn from both, or does an intrinsic mechanism allow our brains to learn the complementary features by perceiving only the

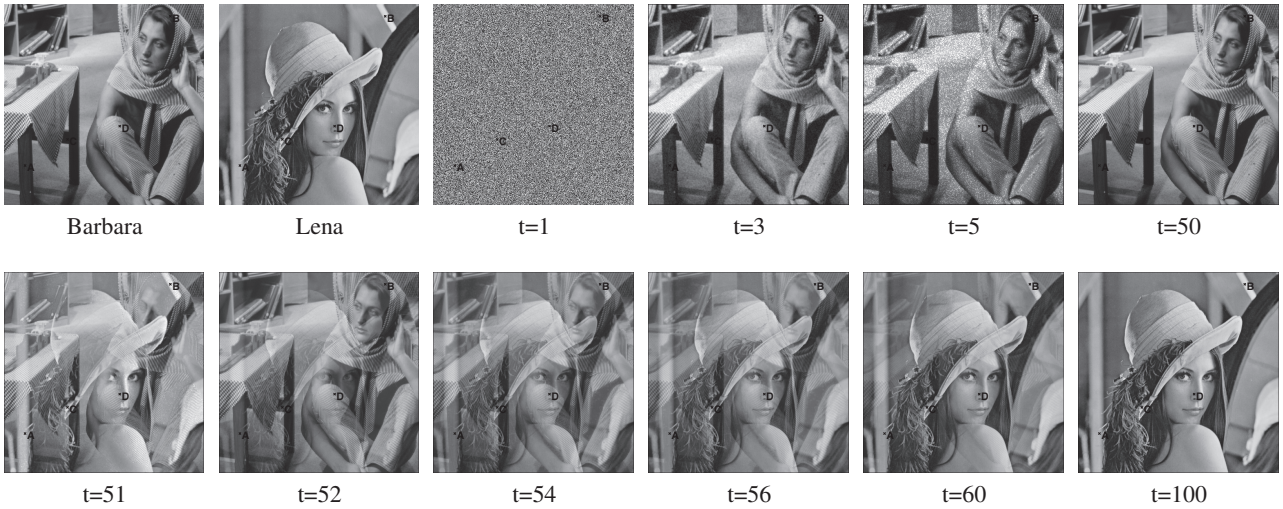



Figure 3: Learning from surprise by the input layer of our model. Emergence of iconic memory is noteworthy. Reproduced from (Banerjee and Dutta 2013c).

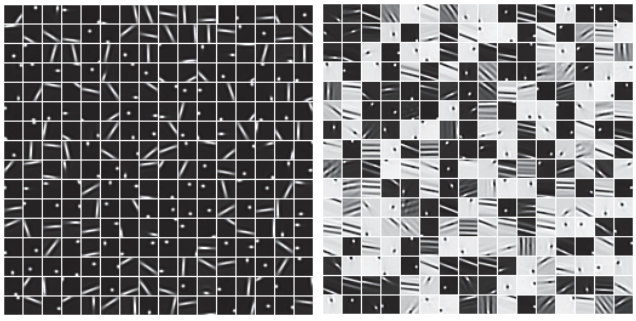



Figure 4: 256 features learned in simple layer from natural images with non-negative (a) and unconstrained (b) weights. Reproduced from (Banerjee and Dutta 2013c).

instances? Consider this higher level example: *All birds Tom has seen can fly. One day he sees a car run over a bird. While trying to explain this incident, he realizes it can be best explained by assuming some birds cannot fly. So, even though he has never seen a bird that could not fly, he learned that fact by explaining some input.* We are not aware of any experiment in the literature that sheds light on this issue.

In conclusion, we presented a multilayered neural architecture learned by the SELP cycles. The feedforward connections to simple and complex neurons encode spatial and temporal sets which correspond to spatiotemporal objects (nouns) and their transformations (adjectives) respectively. Such a transformation over time, encoded by the lateral connections within a simple node in conjunction with the feedforward connections to a complex layer, may be conceived as a mental action on the object it is associated with, thereby performing the function of a verb. The spatial and temporal transformations of these actions, learned by the higher layer

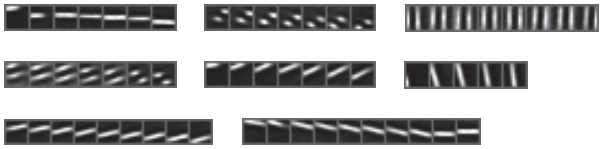


Figure 5: Eight transformations learned by complex neurons from natural videos. All weights were constrained to be non-negative. Reproduced from (Dutta and Banerjee 2013).

neurons, correspond to adverbs. Lateral connections within a node and across nodes associate objects temporally and spatially respectively, corresponding to prepositions.

## Acknowledgment

This research was supported by NSF CISE Grant 1231620.

## References

Bach-y-Rita, P., and Kercel, S. W. 2003. Sensory substitution and the human-machine interface. *Trends in Cognitive Sci.* 7(12):541–546.

Banerjee, B., and Dutta, J. K. 2013a. Efficient learning from explanation of prediction errors in streaming data. In *IEEE BigData Workshop on Scalable Machine Learning*, [Forthcoming].

Banerjee, B., and Dutta, J. K. 2013b. Hierarchical feature learning from sensorial data by spherical clustering. In *IEEE BigData Workshop on Scalable Machine Learning*, [Forthcoming].

Banerjee, B., and Dutta, J. K. 2013c. SELP: A general-purpose framework for learning the norms from saliencies in spatiotemporal data. *Neurocomputing: Special Issue on Brain Inspired Models of Cognitive Memory* [Forthcoming].

Bar, M., et al. 2006. Top-down facilitation of visual recognition. *Proc. Natl. Acad. Sci.* 103(2):449–454.

Bar, M. 2007. The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sci.* 11(7):280–289.

Betsch, B. Y.; Einhäuser, W.; Körding, K. P.; and König, P. 2004. The world from a cat's perspective – statistics of natural videos. *Biol. Cybernetics* 90(1):41–50.

Chalasani, R., and Principe, J. C. 2013. Deep predictive coding networks. *Comput. Res. Repos.* arXiv:1301.3541.

Constantine-Paton, M., and Law, M. I. 1978. Eye-specific termination bands in tecta of three-eyed frogs. *Science* 202(4368):639–641.

David, S. V.; Mesgarani, N.; Fritz, J. B.; and Shamma, S. A. 2009. Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. *J. Neurosci.* 29(11):3374–3386.

Denham, S. L., and Winkler, I. 2006. The role of predictive models in the formation of auditory streams. *J. Physiology – Paris* 100(1):154–170.

Douglas, R. J., and Martin, K. A. C. 2010. *Canonical cortical circuits*. Handbook of Brain Microcircuits. 15–21.

Dutta, J. K., and Banerjee, B. 2013. Learning features and their transformations by spatial and temporal spherical clustering. *Comput. Res. Repos.* arXiv:1308.2350.

Egner, T.; Monti, J. M.; and Summerfield, C. 2010. Expectation and surprise determine neural population responses in the ventral visual stream. *J. Neurosci.* 30(49):16601–16608.

Fenske, M. J.; Aminoff, E.; Gronau, N.; and Bar, M. 2006. Top-down facilitation of visual object recognition: Object-based and context-based contributions. *Progress in Brain Res.* 155:3–21.

Fiorillo, C. D. 2008. Towards a general theory of neural computation based on prediction by single neurons. *PLoS ONE* 3(10).

Friston, K. J. 2008. Hierarchical models in the brain. *PLoS Comput. Biology* 4(11):e1000211.

Fukushima, K. 2003. Neocognitron for handwritten digit recognition. *Neurocomputing* 51(1):161–180.

Gilbert, C. D., and Wiesel, T. N. 1983. Clustered intrinsic connections in cat visual cortex. *J. Neurosci.* 3:1116–1133.

Gilbert, C. D., and Wiesel, T. N. 1989. Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J. Neurosci.* 9(7):2432–2442.

Gill, P.; Woolley, S. M. N.; Fremouw, T.; and Theunissen, F. E. 2008. What's that sound? Auditory area CLM encodes stimulus surprise, not intensity or intensity changes. *J. Neurophysiology* 99:2809–2820.

Hesselmann, G.; Sadaghiani, S.; Friston, K. J.; and Kleinschmidt, A. 2010. Predictive coding or evidence accumulation? False inference and neuronal fluctuations. *PLoS ONE* 5(3):e9926.

Hirsch, J. A., and Gilbert, C. D. 1991. Synaptic physiology of horizontal connections in the cat's visual cortex. *J. Neurosci.* 11(6):1800–1809.

Hosoya, T.; Baccus, S. A.; and Meister, M. 2005. Dynamic predictive coding by the retina. *Nature* 436:71–77.

Jehee, J. F.; Rothkopf, C.; Beck, J. M.; and Ballard, D. H. 2006. Learning receptive fields using predictive feedback. *J. Physiology – Paris* 100(1-3):125–132.

Kilner, J. M.; Friston, K. J.; and Frith, C. D. 2007. Predictive coding: An account of the mirror neuron system. *Cognitive Processing* 8(3):159–166.

Kouh, M., and Poggio, T. 2008. A canonical neural circuit for cortical nonlinear operations. *Neural Computation* 20:1427–1451.

LeCun, Y., and Bengio, Y. 1995. Convolutional networks for images, speech and time series. In Arbib, M. A., ed., *The Handbook of Brain Theory and Neural Networks*. MIT Press. 255–258.

Lee, T. S., and Mumford, D. 2003. Hierarchical Bayesian inference in the visual cortex. *J. Optical Society of America* 20(7):1434–1448.

Matsubara, J. A.; Cynader, M. S.; and Swindale, N. V. 1987. Anatomical properties and physiological correlates of the intrinsic connections in cat area 18. *J. Neurosci.* 7:1428–1446.

Metin, C., and Frost, D. O. 1989. Visual responses of neurons in somatosensory cortex of hamsters with experimentally induced retinal projections to somatosensory thalamus. *Proc. Natl. Acad. Sci.* 86(1):357–361.

Meyer, T., and Olson, C. R. 2011. Statistical learning of visual transitions in monkey inferotemporal cortex. *Proc. Natl. Acad. Sci.* 108(48):19401–19406.

Pati, Y. C.; Rezaiifar, R.; and Krishnaprasad, P. S. 1993. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *27th Asilomar Conf. Signals, Systems and Computers*, 40–44. IEEE.

Rao, R. P. N., and Ballard, D. H. 1999. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neurosci.* 2:79–87.

Rao, R. P. N. 1999. An optimal estimation approach to visual perception and learning. *Vision Res.* 39(11):1963–1989.

Rockland, K. S., and Lund, J. S. 1983. Intrinsic laminar lattice connections in primate visual cortex. *J. Comparative Neurology* 216:303–318.

Rust, N.; Schwartz, O.; Movshon, J. A.; and Simoncelli, E. P. 2005. Spatiotemporal elements of macaque V1 receptive fields. *Neuron* 6(6):945–956.

Samsonovich, A. V. 2010. Toward a unified catalog of implemented cognitive architectures. In Samsonovich, A. V.; Jóhannsdóttir, K. R.; Chella, A.; and Goertzel, B., eds., *Biologically Inspired Cognitive Architectures 2010: Proc. First Annual Meeting of the BICA Society*, 195–244. Amsterdam, Netherlands: IOS Press.

Serre, T.; Wolf, L.; Bileschi, S.; Riesenhuber, M.; and Poggio, T. 2007. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Analysis and Machine Intelligence* 29:411–426.

Spratling, M. W. 2011. A single functional model accounts for the distinct properties of suppression in cortical area V1. *Vision Res.* 51(6):563–576.

Srinivasan, M. V.; Laughlin, S. B.; and Dubs, A. 1982. Predictive coding: A fresh view of inhibition in the retina. *Proc. Royal Society B: Biol. Sci.* 216(1205):427–459.

van Wassenhove, V.; Grant, K. W.; and Poeppel, D. 2005. Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci.* 102(4):1181–1186.

von Melchner, L.; Pallas, S. L.; and Sur, M. 2000. Visual behaviour mediated by retinal projections directed to the auditory pathway. *Nature* 404(6780):871–876.

Vuust, P.; Ostergaard, L.; Pallesen, K. J.; Bailey, C.; and Roepstorff, A. 2009. Predictive coding of music – Brain responses to rhythmic incongruity. *Cortex* 45(1):80–92.

Wacongne, C.; Changeux, J. P.; and Dehaene, S. 2012. A neuronal model of predictive coding accounting for the mismatch negativity. *J. Neurosci.* 32(11):3665–3678.

Winkler, I.; Denham, S.; Mill, R.; Bőhm, T. M.; and Bendixen, A. 2012. Multistability in auditory stream segregation: A predictive coding view. *Phil. Trans. Royal Society B: Biol. Sci.* 367(1591):1001–1012.