Six principles for biologically based computational models of cortical cognition

Randall C. O'Reilly

This review describes and motivates six principles for computational cognitive neuroscience models: biological realism, distributed representations, inhibitory competition, bidirectional activation propagation, error-driven task learning, and Hebbian model learning. Although these principles are supported by a number of cognitive, computational and biological motivations, the prototypical neural-network model (a feedforward back-propagation network) incorporates only two of them, and no widely used model incorporates all of them. It is argued here that these principles should be integrated into a coherent overall framework, and some potential synergies and conflicts in doing so are discussed.

A number of important principles have been developed for computational neural-network models of cortical learning and cognitive processing. However, relatively little work has been done to try to integrate these principles into a coherent overall framework. Integrating these principles allows one to demonstrate the consistency of different models, capitalize on synergies between different principles, organize and consolidate existing findings, and generate novel insights into the nature of cognition. This review describes and motivates a provisional set of six principles (illustrated in Fig. 1) that have proven individually useful in a number of existing models. Despite their proven utility, most models incorporate only a small number of these principles (e.g. the prototypical feedforward back-propagation network uses only two). Thus, this review attempts to highlight the potential advantages and pitfalls of using a more inclusive set of principles.

Although a specific algorithm called Leabra has been developed to incorporate these principles (see Box 1), this review focuses on the history and importance of the principles themselves, and the ways in which these principles interact with each other. As an important caveat, this discussion focuses on biologically based principles that are particularly relevant for cognition, and does not include a number of functional and cognitive-level principles that could also be enumerated.

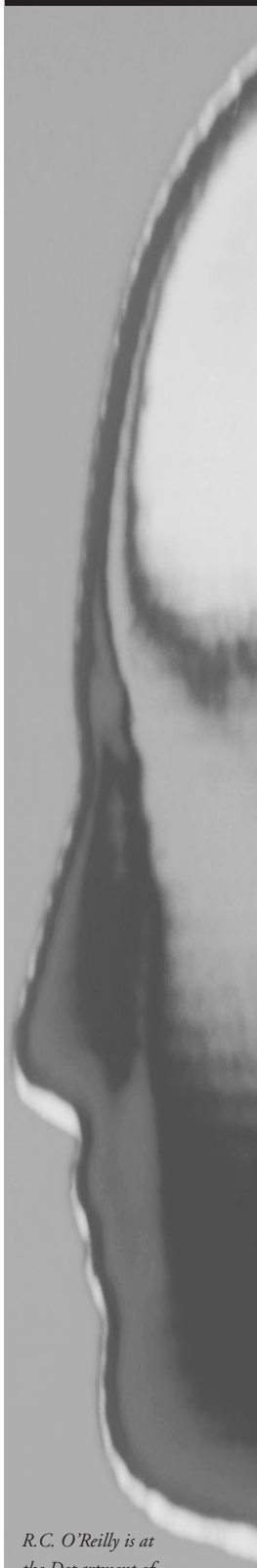
The proposed set of principles can be considered an extension of the 'GRAIN' framework of McClelland[1]. GRAIN stands for graded, random, adaptive, interactive, (nonlinear) network. This framework was primarily motivated by (and applied to) issues surrounding the dynamics of activation flow through a neural network. By way of extension, the present framework emphasizes learning mechanisms and the architectural properties that support them. Two of the key principles in GRAIN, interactivity and competition, are among the six principles emphasized here. The other GRAIN principles (graded, nonlinear activations, graded activation propagation, and intrinsic variability) are assumed, but not emphasized in this framework because of their nearly ubiquitous acceptance within neural-network models (but see Ref. 2 for an interesting application of these principles to controversies in cognitive development).

The principles

The six principles can be grouped into three categories. The first principle, biological realism, is in a category by itself, providing a general overriding constraint on the framework. The next three principles, distributed representations, inhibitory competition, and bidirectional activation propagation (interactivity), are concerned with the architecture of the network and the general behavior of the neuron-like processing units within it. The final two principles, error-driven task learning and Hebbian model learning, govern the way that learning occurs in the network.

(1) Biological realism

Biological realism lies at the foundation of the entire enterprise of computational modeling in cognitive neuroscience. This approach seeks to understand how the brain (and specifically the cortex in the present case) gives rise to cognition, not how some abstraction of uncertain validity does so. Thus, wherever possible, computational models should



R.C. O'Reilly is at the Department of Psychology, University of Colorado at Boulder, Campus Box 345, Boulder, CO 80304, USA.

tel: +1 303 492 0054
fax: +1 303 492 2967
e-mail: oreilly@psych.colorado.edu

1364-6613/98/$ – see front matter PII: S1364-6613(98)01241-8

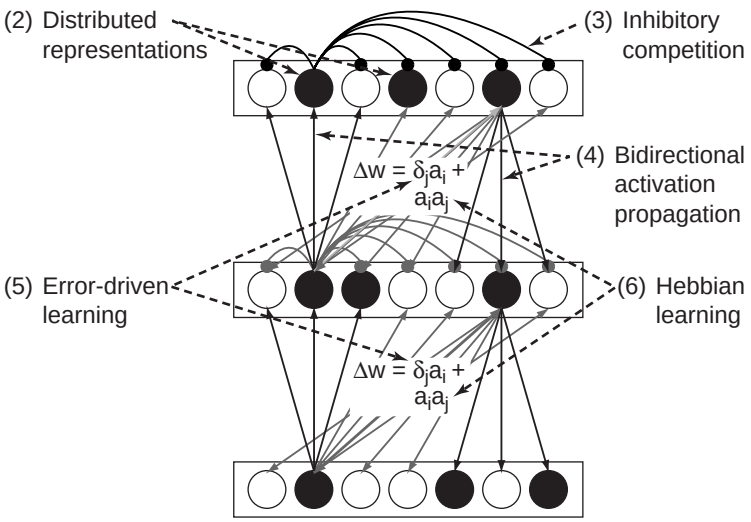


Fig. 1 Illustration of the six core principles, and their instantiation in a neural network. Biological realism (1) is an overarching constraint. Distributed representations (2) have multiple units active, while inhibitory competition (3), (implemented via inhibitory connectivity) ensures that relatively few such units are active. Bidirectional activation propagation (4), (implemented by bidirectional connectivity) enables both bottom-up and top-down constraints to simultaneously shape the internal representation. Error-driven learning (5) shapes representations according to differences between expected outputs and actual ones (represented by the error term δ_j). Hebbian learning (6) shapes representations according to the co-occurrence (correlation) statistics of items in the environment (represented by the product of the sending- and receiving-unit activations).

be constrained and informed by biological properties of the cortex. Moreover, computational mechanisms that violate known biological properties should not be relied upon. This point has implications for error-driven learning, as discussed below.

Although the issue of biological realism is easy to state, it can be difficult to apply, because the known biology often does not provide sufficient constraints. Thus, biological realism often reduces to plausibility arguments, which depend on things like how simple and local the mechanism in question is, and that it is not inconsistent with known biology. Also, one can adopt a converging evidence approach, where multiple constraints from biology, computation, and cognition converge to support a given principle. This approach is emphasized here.

Architectural principles

(2) Distributed representations

The cortex is widely believed to use distributed representations to encode information. A distributed representation uses multiple active neuron-like processing units to encode information (as opposed to a single unit, localist representation), and the same unit can participate in multiple representations. Each unit in a distributed representation can be thought of as representing a single feature, with information being encoded by particular combinations of such features. Electrophysiological recordings demonstrate that distributed representations are widely used in the cortex (e.g. Refs 3–5). The functional benefits of distributed representations include greater efficiency, robustness, and accuracy, and the ability to represent similarity relationships[6]. The efficiency of distributed representations can be appreciated by analogy with letters. Just as different combinations of a small number of letters can represent a large number of words, so can different combinations of a small set of units represent a large amount of information. The robustness of distributed representations comes from the redundancy of having each item represented by many units. Distributed representations can more accurately represent graded values through coarse coding, where a value is encoded by the relative magnitudes of a number of broadly tuned units. Finally, similarity is represented by the shared units involved in the distributed representations of different items.

(3) Inhibitory competition

Inhibitory competition is important because it selects representations for processing and for subsequent refinement over learning. Inhibitory competition arises when mutual inhibition among a set of units (i.e. as mediated by inhibitory interneurons) prevents all but a subset of them from becoming active at a time. Approximately 20% of the neurons in the cortex are inhibitory interneurons[7], and it is clear that they control the explosion of activation that would otherwise result from all the positive interconnectivity among cortical pyramidal neurons (e.g. as happens in epilepsy). Inhibitory competition allows only the most strongly excited representations to prevail, with this selection process identifying the most appropriate representations for subsequent processing. Furthermore, most learning mechanisms (including those discussed later) are affected by this selection process such that only the selected representations are refined over time through learning, resulting in an effective differentiation and distribution of representations[8–11].

Aside from the selection and refinement of representations, another benefit of inhibitory competition comes from the idea that, given the general structure of the environment, sparse distributed representations (i.e. having relatively few units active at a time) are particularly useful[12,13]. For example, in visual processing, a given object can be defined along a set of feature dimensions (e.g. shape, size, color, texture), with a large number of different values along each dimension (i.e. many different possible shapes, sizes, colors, textures, etc.). Assuming that the individual units in a distributed representation encode these feature values, a representation of a given object will only activate a small subset of units (i.e. the representations will be sparse). To further substantiate this argument, Olshausen and Field[14] showed that imposing a bias for developing sparse distributed representations can result in the development of realistic early visual representations (oriented edge detectors) of natural visual scenes. More generally, it seems as though the world can be usefully represented in terms of a large number of categories with a large number of exemplars per category (animals, furniture, trees, etc.). If we again assume that only a relatively few such exemplars are processed at a given time, a bias favoring sparse representations is appropriate.

(4) Bidirectional activation propagation (interactivity)

Bidirectional activation propagation is a critical principle for information flow through the network. Bidirectional

activation propagation (also called 'interactivity' or 'recurrence') is the communication of activation simultaneously in both bottom-up and top-down directions. This contrasts with feedforward activation propagation where information only goes in one direction (bottom-up). To enable information to flow in both directions simultaneously in a stable and effective manner, processing must proceed in gradual, iterative steps. Thus, a temporally-extended 'settling' process with many iterative steps is required for the network to achieve an appropriate representation of a given input pattern. This is a central feature of GRAIN (Ref. 1). Bidirectional connectivity is ubiquitous in the cortex (e.g. Refs 15–17). An important benefit of bidirectional activation propagation is powerful 'constraint-satisfaction' processing[18,19], where both lower-level (e.g. perceptual) and higher-level (e.g. conceptual) constraints can be simultaneously brought to bear in interpreting and processing inputs.

The importance of interactivity for understanding cognitive processing was demonstrated in the word superiority model of McClelland and Rumelhart[9]. They showed that interactivity could explain the counterintuitive finding that higher-level word processing can influence lower-level letter perception. More recently, Vecera and O'Reilly[20] showed that bidirectional constraint satisfaction can model people's ability to resolve ambiguous visual inputs in favor of familiar versus novel objects[21]. They also showed that adding inhibitory competition to an interactive network significantly speeded the settling process, and greatly reduced the number of times the network settled into bad local minima.

Learning principles

Learning is essential for shaping the representations of neural networks according to the structure of the environment. A key issue is what aspects of the environmental structure should be learned, with the understanding that not everything can or should be represented. The following two learning principles exploit two complementary aspects of environmental structure: task demands, and the extent to which different things co-occur. The first is referred to as 'task learning' for obvious reasons, and the second is referred to as 'model learning' because the objective is to develop an internal model of the environment irrespective of specific tasks. These two learning objectives can be achieved by two different forms of implementational mechanisms, 'error-driven' and 'Hebbian' learning, respectively.

(5) Error-driven task learning

Error-driven learning (also called 'supervised' learning) is important for shaping representations according to task demands by learning to minimize the difference (i.e. the error) between a desired outcome and what the network actually produced. This principle captures the idea that you learn what enables you to succeed at the necessary tasks of life, without bothering to represent aspects of the environment that are not relevant to these tasks. The widely used back-propagation learning algorithm[22] directly minimizes error through gradient descent, and has proven to be very powerful. Although task learning is clearly psychologically relevant, and a majority of psychological models have used this form of learning, its biological plausibility has been widely questioned because it requires the propagation of error signals in a manner inconsistent with known neurobiological properties (e.g. Refs 23,24). Furthermore, it has not been clear where the necessary 'teaching' signals could plausibly come from. However, it has recently been shown that biologically plausible bidirectional activation propagation (see principle 4) can be used to perform essentially the same error-driven learning as backpropagation[25], using any of a number of readily available teaching signals. The resulting algorithm generalizes the recirculation algorithm of Hinton and McClelland[26], and is thus called 'GeneRec'.

The basic idea behind GeneRec is that instead of propagating an error signal, which is a difference between two terms, one can propagate the two terms separately as activation signals, and then take their difference locally at each unit. This works by having two phases of activations for computing the two terms. In the 'expectation' phase, the bidirectionally-connected network settles based on an input activation pattern into a state that reflects the expected consequences or correlates of that input pattern. Then, in the 'outcome' phase, the network experiences actual consequence(s) or correlate(s). The difference between outcome and expectation is the error signal, and the bidirectional connectivity propagates this error signal throughout the network via local activation signals. Thus, interactivity enables units everywhere in the network to receive (possibly indirectly via hidden layers) activation signals from the layer(s) where the expectation and outcome are represented. The remarkable thing is that the activation signals in an interactive network are naturally propagated (even through hidden layers) in just the right way to enable the correct error gradient to be simply and locally computed at each unit[25].

The GeneRec analysis also showed that Boltzmann machine learning and its deterministic versions[19,27–29] can be seen as variants of this more biologically plausible version of the back-propagation algorithm. This means that all of the existing approaches to error-driven learning using activation-based signals converge on essentially the same basic mechanism, making it more plausible that this is the way the brain does error-driven learning. Furthermore, the form of synaptic modification necessary to implement this algorithm is consistent with (though not directly validated by) known properties of biological synaptic modification mechanisms. Finally, there are many sources in the natural environment for the necessary outcome phase signals in the form of actual environmental outcomes that can be compared with internal expectations to provide error signals[25,30]. Thus, one does not need to have an explicit 'teacher' to perform error-driven learning. Taken together, these developments make it difficult to continue to object to the use of error-driven learning on the grounds that it is not biologically plausible.

(6) Hebbian model learning

Model learning (also called self-organizing or 'unsupervised' learning) is important for forming internal representations of the general (statistical) structure of the environment, without respect to particular tasks. Many versions of this general idea exist, defined by what aspects of environmental structure are

Box 1. The Leabra implementation

The six principles have been implemented in an algorithm called Leabra, which is briefly presented here. Leabra stands for 'learning in an error-driven and associative, biologically realistic algorithm' (where associative is another term for Hebbian learning). Leabra has been used in a forthcoming textbook[a] to implement a wide range of cognitive neuroscience models. The scope of phenomena it is capable of modeling is commensurate with the breadth of the principles as discussed in the paper, and demonstrates their sufficiency and mutual compatibility.

Point-neuron activation function

Leabra uses a point-neuron activation function that models the electrophysiological properties of real neurons, while simplifying their geometry to a single point. This is nearly as simple computationally as the standard sigmoidal activation function, but the more biologically based implementation makes it considerably easier to model inhibitory competition, as described below. Further, it enables cognitive models to be more easily related to more physiologically detailed simulations, thereby facilitating bridge-building between biology and cognition.

The membrane potential V_m is updated as a function of ionic conductances g with reversal (driving) potentials E as follows:

$$\frac{dV_m(t)}{dt} = \tau \sum_c g_c(t)\overline{g_c}(E_c - V_m(t))$$

with three channels, c, corresponding to: e, excitatory input; l, leak current; and i, inhibitory input. The equilibrium potential can be written in a simplified form by setting the excitatory driving potential, E_e, to one and the leak and inhibitory driving potentials, E_l and E_i, to zero:

$$V_m^{\infty} = \frac{\overline{g_e} g_e}{\overline{g_e} g_e + \overline{g_i} g_i + \overline{g_l} g_l}$$

This shows that the neuron is computing a balance between excitation and the opposing forces of leak and inhibition. This form of the equation can be understood in terms of a Bayesian decision-making framework[a]. Activation communicated to other cells (y) is a thresholded (Θ), sigmoidal function of the membrane potential with gain parameter γ:

$$y_j(t) = \frac{1}{\left(1 + \frac{1}{\gamma [V_m(t) - \Theta]_+}\right)}$$

This can be convolved with Gaussian noise, producing a less discontinuous function with a softer lower threshold.

***k*-Winners-take-all inhibition (KWTA)**

Leabra uses a KWTA function to achieve sparse distributed representations. This function is achieved by setting a uniform level of inhibition for all units in the layer that prevents more than k units from getting over threshold. This inhibitory current is given by:

$$g_i = g^{\Theta}_{k+1} + q(g^{\Theta}_k - g^{\Theta}_{k+1})$$

where q is typically 0.25, and the threshold-level inhibition terms are

$$g^{\Theta} = \frac{\sum_{c \neq i} g_c \overline{g_c}(E_c - \Theta)}{\Theta - E_i}$$

for the units with the kth and $k+1$th highest excitatory inputs. Activation dynamics similar to those produced by this function have been shown to result from simulated inhibitory interneurons that project both feedforward and feedback inhibition[a].

Error-driven learning

Error-driven learning is implemented in Leabra using a symmetric version of the biologically plausible GeneRec algorithm[b], that is functionally equivalent to the deterministic Boltzmann machine and contrastive Hebbian learning (CHL) (Refs c,d). The network settles in two phases, an expectation (minus) phase and an outcome (plus) phase, and then computes a simple difference of a pre- and postsynaptic activation product across these two phases:

$$\Delta w_{ij} = x^+_i y^+_j - x^-_i y^-_j$$

for sending unit x_i, and receiving unit y_j, in the two phases.

Hebbian learning

The simplest form of Hebbian learning adjusts the weights in proportion to the product of the sending (x_i) and receiving (y_j) unit activations:

$$\Delta w_{ij} = x_i y_j$$

The weight vector is dominated by the principal eigenvector of the pairwise correlation matrix of the input, but it also grows without bound. Leabra uses a variant of the Oja normalization[e]:

$$\Delta w_{ij} = x_i y_j - y_j w_{ij} = y_j (x_i - w_{ij})$$

which can also be seen as computing the expected value of the sending unit activity conditional on the receiver's activity, if treated like a binary variable active with probability y_j:

$$w_{ij} \approx \langle x_i | y_i \rangle_p$$

This is essentially the same rule as used in standard competitive learning or mixtures-of-Gaussians[f,g].

Error-driven and Hebbian learning are linearly combined at each synapse in the network, using a normalized mixing constant. To keep the error-driven component within the same 0–1 range of the Hebbian term, soft-weight bounding with exponential approach to these extremes is used on this component. Finally, a sigmoidal contrast-enhancement function on the weights can be used to facilitate learning in environments with underlying binary features (i.e. imposing a bias towards binary weights). (See Ref. 1 for details.)

Principal results

In O'Reilly *et al.*[a], Leabra is used to replicate a large number of published models that were originally implemented using a variety of different algorithms from backpropagation to Hebbian self-organizing learning. Leabra also illustrates many of the issues

deemed important to represent, but it is generally agreed that something like correlational structure is important. Hebbian learning mechanisms represent this correlational structure, encoding the extent to which different things co-occur in the environment[31]. Biologically, Hebbian learning requires that the synaptic strength change as a function of the co-activation of the sending and receiving neurons. NMDA-mediated long-term potentiation (LTP) has this Hebbian property (e.g. Ref. 32). Thus, Hebbian learning is almost universally regarded as being biologically plausible. At a functional level, the co-occurrence of items suggests that there might be a causal relationship between them.

discussed in this paper regarding the interactions among the different principles. For example, just adding interactivity to an otherwise generic error-driven network (e.g. a GeneRec network) significantly impairs generalization performance. However, also adding Hebbian learning and inhibitory competition (in Leabra) restores good generalization performance within an interactive network (R.C. O'Reilly, PhD thesis, Carnegie Mellon University, 1996). The conclusion is similar to that of the GRAIN exploration of interactivity and noise – interactivity itself may cause problems, but these can be remedied with additional principles.

In addition to replicating existing models, Leabra also provides better models of several phenomena. One salient example of this is in the case of the U-shaped past-tense over-regularization phenomenon, which has proven difficult to capture using purely error-driven backpropagation networks without also manipulating the training environment in a potentially implausible fashion[h–j]. By adding Hebbian learning and inhibitory competition, Leabra introduces biases that produce a much more pronounced U-shaped effect (including a longer period of early competence)[a]. This can be contrasted with the essentially monotonic decrease in over-regularizations that, in retrospect, is exactly what would be expected from a purely error-driven algorithm (see Ref. a and Hoeffner, PhD Thesis, Carnegie Mellon University, 1997, for a more detailed critique of existing models).

References

a O'Reilly, R.C., Munakata, Y. and McClelland, J.L. *Explorations in Computational Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*, MIT Press (in press)

b O'Reilly, R.C. (1996) Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm *Neural Comput.* 8, 895–938

c Hinton, G.E. (1989) Deterministic Boltzmann learning performs steepest descent in weight-space *Neural Comput.* 1, 143–150

d Movellan, J.R. (1990) Contrastive Hebbian learning in the continuous Hopfield model, in *Proceedings of the 1989 Connectionist Models Summer School* (Touretzky, D.S., Hinton, G.E. and Sejnowski, T.J., eds), pp. 10–17, Morgan Kaufman

e Oja, E. (1982) A simplified neuron model as a principal component analyser *J. Math. Biol.* 15, 267–273

f Rumelhart, D.E. and Zipser, D. (1986) Feature discovery by competitive learning, in *Parallel Distributed Processing (Vol. 1): Foundations* (Rumelhart, D.E., McClelland, J.L. and PDP Research Group, eds), pp. 151–193, MIT Press

g Nowlan, S.J. (1990) Maximum likelihood competitive learning, in *Advances in Neural Information Processing Systems* (Vol. 2) (Touretzky, D.S., ed.), pp. 574–582, Morgan Kaufman

h Rumelhart, D.E. and McClelland, J.L. (1986) On learning the past tenses of English verbs, in *Parallel Distributed Processing (Vol. 2): Psychological and Biological Models* (Rumelhart, D.E., McClelland, J.L. and PDP Research Group, eds), pp. 216–271, MIT Press

i Plunkett, K. and Marchman, V.A. (1993) From rote learning to system building: acquiring verb morphology in children and connectionist nets *Cognition* 48, 21–69

j Marcus, G.F. *et al.* (1992) Over-regularization in language acquisition *Monogr. Soc. Res. Child Dev.* 57, 1–165

Furthermore, co-occurring items can be more efficiently represented together within a common representational structure. Mathematical analyses have shown that Hebbian learning performs something like principal components analysis[33], which extracts the principal dimensions of covariance within the environment. An interesting demonstration of the power of this kind of Hebbian model learning was recently provided in the form of a model that performs principal components analysis on the co-occurrence statistics of words within large texts, yielding surprisingly powerful representations of word meaning[34].

Interactions among the principles

The preceding discussion provided specific and compelling motivations for each of the individual principles. In this section, three examples of interactions (synergies and conflicts) among the six principles will be discussed. The first example comes from the GRAIN framework, and deals with the consequences of interactivity and noise. The second explores the interactions between distributed representations and competition, which can be at odds with each other. The last explores the interactions between error-driven and Hebbian learning.

Interactivity and noise

The GRAIN framework has been used to explore the implications of some of the principles on the activation dynamics of a network[1]. For example, although interactivity (bidirectional activation propagation) is important for accounting for a range of different behavioral phenomena, it can also be problematic for others. Specifically, interactivity interfered with the ability of a network to exhibit independent contributions from context and stimulus strength in a stimulus identification situation[35,36]. McClelland showed that the use of intrinsic variability (noise) can resolve this conflict, resulting in a model that captures a wider range of phenomena, including standard interactive phenomena (e.g. top-down effects) and the independent contributions of context and stimuli[1].

This example illustrates a theme that emerges repeatedly when attempting to integrate different principles (see Box 1 for another example): often, subsets of principles do not work as well as a more complete set of principles. Thus, instead of abandoning a principle (e.g. interactivity) when it appears to introduce a problem, one should consider how other principles might be adopted that would resolve the problem. The advantage of the integrative approach is that the resulting model then accounts for a much wider range of phenomena, and may provide important new insights into the nature of the originally problematic phenomenon. For example, the GRAIN model can explain the conditions under which you would not expect to find an independent contribution of stimulus and context (see Ref. 1 for details and empirical validation).

Distributed representations and competition

Perhaps one of the most important challenges in integrating the six principles comes in combining distributed representations and competition, which tend to work at cross purposes. Distributed representations require multiple active units that cooperatively represent something, whereas competition causes fewer units to be active, and it can inhibit cooperativity. A reasonable compromise between these two principles is the sparse distributed representation as discussed previously. Although seemingly straightforward, achieving a sparse distributed representation is technically challenging, primarily because this case is difficult to analyze

mathematically within a probabilistic learning framework. The problem is one of combinatorial explosion – one needs to take into account all the different possible combinations of active and inactive units to analyze a sparse distributed representation based on true inhibitory competition. Thus, sparse distributed representations fall in a complex intermediate zone between two easily analyzed frameworks[37]: (i) The winner-take-all (WTA) framework[10,11,38], where only one unit is allowed to be active at a time. Having a single active unit eliminates the combinatorial problems, but this also violates principle 2 by not allowing for distributed representations. (ii) The independent units framework, where the units are considered to be (conditionally) independent of each other (e.g. a standard back-propagation network). This allows the combined probability of an activation pattern to be represented as a simple product of the individual unit probabilities (and for distributed representations), but it also violates principle 3 because there is no competition.

There have been a number of attempts to remedy the limitations of these two analytical frameworks, by introducing distributed representations within a basically WTA framework, or by introducing sparseness constraints within the independent units framework. However, the basic limitations of these frameworks are difficult to overcome. Basically, any use of WTA prevents the cooperativity and combinatoriality of true distributed representations, and the need to preserve independence among the units in the independent units framework prevents the introduction of any true activation-based competition. After discussing these approaches and their limitations, the more difficult to analyze approach of directly implementing sparse distributed representations using inhibitory competition will be discussed.

The following are extensions of the WTA framework. In the mixture-of-experts framework[39], a WTA competition takes place within a specialized 'gating' network that regulates the participation of a set of 'expert' networks, which can themselves have distributed representations. A limitation of this approach is the coarse-grained level of the competition – whole groups of units compete, but individual units do not. Also, multiple experts cannot easily cooperate due to the WTA limitation. The model of Dayan and Zemel[40] uses a WTA assumption where units in the hidden layer compete to determine the value of a given unit in the output. However, this simply transfers the WTA assumption from representing the input to representing the output, and a WTA assumption anywhere is likely to be problematic. The Dayan and Zemel model was intended as an improvement over the 'noisy-or' model of Saund[41], which did not result in sufficient competition. Finally, the Kohonen network[8] uses a WTA to select a single winner, but then a neighborhood of units around that winner are also activated. Although useful for achieving topographic representations, this kind of fixed, imposed activation state does not enable the full combinatorial representational power that is an essential feature of true distributed representations.

Within the independent units framework, the main approach has been to introduce a sparseness constraint that does not actually involve direct activation-based competition. This usually involves adding an extra factor to the learning rule that favors sparse representations (e.g. R.S. Zemel, PhD Thesis, University of Toronto, 1993; and Refs 14,42,43), or adding a sparseness bias into the activation function itself (e.g. Ref. 37). Thus, units are only competing over the long time-course of learning (or against their own negative bias), and not directly with one another to represent the current input pattern (i.e. selection). Furthermore, the dynamic thresholding behavior one achieves with activation-based competition (which, for example, makes the system robust to changes in absolute levels of excitation) is not present in these approaches. This limitation is particularly evident in bidirectionally connected networks, where the need to control positive feedback requires the dynamic thresholding of true competition (R.C. O'Reilly, PhD Thesis, Carnegie Mellon University, 1996). Thus, integrating all of the principles places further demands on the competition mechanism.

It seems clear that the cortex implements inhibitory competition (and sparse distributed representations) via inhibitory interneurons. One way of understanding the effects of this inhibitory competition is in terms of a '*k*-winners-take-all' (KWTA) mechanism, which generalizes the WTA approach to *k* winners[44]. A KWTA mechanism can enforce true competition amongst the units, while allowing for a (sparse) distributed representation across the subset of *k* units. KWTA mechanisms have been analyzed for factors such as stability and convergence onto *k* units, and can be implemented with biologically plausible lateral inhibition mechanisms[44,45]. However, they have not been analytically treated within a probabilistic learning framework, due to the combinatorial explosion problems. Nevertheless, a simple form of KWTA that works well in bidirectionally connected networks has been shown to be useful for modeling a wide range of cognitive phenomena (see Box 1).

Learning principles

Before discussing the interactions between error-driven task learning and Hebbian model learning, the distinction between the computational objectives of learning (i.e. task and model learning) and the implementational mechanisms (i.e. error-driven and Hebbian learning) needs to be clarified. Two points of potential confusion exist: (i) error-driven learning can be used to achieve a model-learning objective; and (ii) some error-driven learning mechanisms resemble Hebbian mechanisms. The first point of confusion arises because one can train a network to reproduce the information in the environment using error-driven mechanisms, resulting in a task-independent model of the environment; that is, via an auto-encoder (see Refs 46,47 and R.S. Zemel, PhD thesis, University of Toronto, 1993) or a generative model[14,48]. One can also learn an internal model based on error signals derived from the mismatch between different sensory representations of the same underlying event[49–51]. (This idea can also be viewed as an instance of the GeneRec expectation-outcome framework, where each modality creates an 'expectation' about how the other modality will represent the event. The difference between this expectation and how the modality actually represents the event constitutes the error signal.) These examples raise the issue of why one should use Hebbian mechanisms to implement model

learning, instead of using error-driven learning for both task and model learning. The subsequent discussion of the advantages of combining error-driven and Hebbian learning addresses this issue.

As for the second point of confusion, a version of the GeneRec algorithm[25] is equivalent to the 'contrastive Hebbian learning' (CHL) algorithm of Movellan[29], which uses the difference between two Hebbian terms. Also, other algorithms have been proposed that achieve quasi-error-driven learning with Hebbian-like mechanisms (e.g. Ref. 52). However, despite these apparent similarities in the surface form of the learning rule, error-driven learning achieves a very different computational objective from simple Hebbian learning; only error-driven learning can achieve a fully general, powerful form of task learning (i.e. one that is capable of learning arbitrary input–output mappings).

Thus, it seems clear that we should begin with the assumption that error-driven learning is essential for achieving task learning. From this error-driven perspective, it would then be of interest to know if further constraining the learning with Hebbian model learning would yield any benefits. A general framework for understanding why this might be the case was articulated by Geman, Bienenstock and Doursat[53]. They argued that standard neural networks (e.g. generic back-propagation networks) are typically underconstrained by the learning task, and thus suffer from too much variance in solutions. This can have negative consequences for generalization to novel inputs, among other things. The solution is to add biases to networks that further constrain the learning by favoring particular forms of solutions (representations). To be beneficial, these biases obviously need to be appropriate – there are no generically optimal set of biases – but there may be a set of biases that is particularly appropriate for representing the real world. Indeed, encouraging sparse distributed representations can be seen as just such a bias that has been justified in terms of real world properties, as discussed previously. It is likely, given the general importance of correlational information in the world (e.g. for suggesting causal relationships), that including a Hebbian bias towards representing co-occurrence statistics would be another such generally useful bias.

Although error-driven learning can be sensitive to correlational information, Hebbian learning is directly constrained to learn a correlation-based model because Hebbian weight changes directly reflect unit correlations. Thus, Hebbian model learning can provide a distinct and useful additional bias to further constrain error-driven task learning. This additional Hebbian bias can be thought of as a somewhat 'smarter' version of the widely-used weight decay bias (e.g. Ref. 54). Aside from the Leabra algorithm described in Box 1, there is at least one other example in the literature where error-driven (back-propagation) and Hebbian learning have been combined, with the expected beneficial results[55]. In addition to the synergy between these two forms of learning, combining both task-based and model-based learning enables one to account for phenomena associated specifically with these different types of learning. For example, it seems reasonable to assume that some kinds of learning occur as a result of mere exposure to stimuli (i.e. as would be expected by model learning, but not task learning). However, other kinds of learning (e.g. complex input–output mappings) clearly require task learning.

Outstanding questions

- Are there cognitive phenomena or biological facts that appear to contradict directly the core principles outlined here?
- Is it possible that different parts of the cortex emphasize some principles over others? How might this influence functional specialization in the cortex?
- How many other important principles are missing from the present list?
- Can complex, sequential cognitive processing be shown to emerge from such basic principles as those discussed here, or does this require a whole new set of principles?
- How might error-driven and Hebbian learning co-exist and interact with reinforcement learning, which is likely to be taking place in sub-cortical structures, and possibly the cortex?

Acknowledgements

I thank Yuko Munakata and Jerry Rudy for their helpful comments. This work was supported in part by NIH program project grant MH47566.

References

1 McClelland, J.L. (1993) The GRAIN model: a framework for modeling the dynamics of information processing, in *Attention and Performance (Vol. XIV): Synergies in Experimental Psychology, Artifical Intelligence, and Cognitive Neuroscience* (Meyer, D.E. and Kornblum, S., eds), pp. 655–688, Erlbaum

2 Munakata, Y. *et al.* (1997) Rethinking infant knowledge: toward an adaptive process account of successes and failures in object permanence tasks *Psychol. Rev.* 104, 686–713

3 Desimone, R. and Ungerleider, L.G. (1989) Neural mechanisms of visual processing in monkeys, in *Handbook of Neurophysiology* (Vol. 2) (Boller, F. and Grafman, J., eds), pp. 267–299, Elsevier

4 Georgopoulos, A.P. (1990) Neurophysiology and reaching, in *Attention and Performance* (Vol. 13) (Jeannerod, M., ed.), pp. 227–263, Erlbaum

5 Rao, S.C., Rainer, G. and Miller, E.K. (1997) Integration of what and where in the primate prefrontal cortex *Science* 276, 821–824

6 Hinton, G.E., McClelland, J.L. and Rumelhart, D.E. (1986) Distributed Representations, in *Parallel Distributed Processing (Vol. 1): Foundations* (Rumelhart, D.E., McClelland, J.L. and PDP Research Group, eds), pp. 77–109, MIT Press

7 Gabbot, P.L.A. and Somogyi, P. (1986) Quantitative distribution of GABA–immunoreactive neurons in the visual cortex (area 17) of the cat *Exp. Brain Res.* 61, 323–331

8 Kohonen, T. (1984) *Self–Organization and Associative Memory*, Springer-Verlag

9 McClelland, J.L. and Rumelhart, D.E. (1981) An interactive activation model of context effects in letter perception: 1. An account of basic findings *Psychol. Rev.* 88, 375–407

10 Rumelhart, D.E. and Zipser, D. (1986) Feature discovery by competitive learning, in *Parallel Distributed Processing (Vol. 1): Foundations* (Rumelhart, D.E., McClelland, J.L. and PDP Research Group, eds), pp. 151–193, MIT Press

11 Grossberg, S. (1976) Adaptive pattern classification and universal recoding I: Parallel development and coding of neural feature detectors *Biol. Cybern.* 23, 121–134

12 Barlow, H.B. (1989) Unsupervised learning *Neural Comput.* 1, 295–311

13 Field, D.J. (1994) What is the goal of sensory coding? *Neural Comput.* 6, 559–601

14 Olshausen, B.A. and Field, D.J. (1996) Emergence of simple-cell receptive-field properties by learning a sparse code for natural images *Nature* 381, 607–609

15 Felleman, D.J. and Van Essen, D.C. (1991) Distributed hierarchical processing in the primate cerebral cortex *Cereb. Cortex* 1, 1–47

16 Levitt, J.B. *et al.* (1993) Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 and 46) *J. Comp. Neurol*. 338, 360–376

17 White, E.L. (1989) *Cortical Circuits: Synaptic Organization of the Cerebral Cortex: Structure, Function, and Theory*, Birkhäuser

18 Smolensky, P. (1986) Information processing in dynamical systems: foundations of harmony theory, in *Parallel Distributed Processing (Vol. 1): Foundations* (Rumelhart, D.E., McClelland, J.L. and PDP Research Group, eds), pp. 282–317, MIT Press

19 Ackley, D.H., Hinton, G.E. and Sejnowski, T.J. (1985) A learning algorithm for Boltzmann machines *Cognit. Sci.* 9, 147–169

20 Vecera, S.P. and O'Reilly, R.C. (1998) Figure–ground organization and object recognition processes: an interactive account *J. Exp. Psychol. Hum. Percept. Perform.* 24, 441–462

21 Peterson, M.A. (1994) Object recognition processes can and do operate before figure ground organization *Curr. Dir. Psychol. Sci*. 3, 105–111

22 Rumelhart, D.E, Hinton, G.E. and Williams, R.J. (1986) Learning internal representations by error propagation, in *Parallel Distributed Processing (Vol. 1): Foundations* (Rumelhart, D.E., McClelland, J.L. and PDP Research Group, eds), pp. 318–362, MIT Press

23 Crick, F.H.C. (1989) The recent excitement about neural networks *Nature* 337, 129–132

24 Zipser, D. and Andersen, R.A. (1988) A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons *Nature* 331, 679–684

25 O'Reilly, R.C. (1996) Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm *Neural Comput*. 8, 895–938

26 Hinton, G.E. and McClelland, J.L. (1988) Learning representations by recirculation, in *Neural Information Processing Systems* (Anderson, D.Z., ed.), pp. 358–366, American Institute of Physics

27 Hinton, G.E. (1989) Deterministic Boltzmann learning performs steepest descent in weight-space *Neural Comput.* 1, 143–150

28 Peterson, C. and Anderson, J.R. (1987) A mean-field-theory learning algorithm for neural networks *Complex Syst*. 1, 995–1019

29 Movellan, J.R. (1990) Contrastive Hebbian learning in the continuous Hopfield model, in *Proceedings of the 1989 Connectionist Models Summer School* (Touretzky, D.S, Hinton, G.E. and Sejnowski, T.J., eds), pp. 10–17, Morgan Kaufman

30 McClelland, J.L. (1994) The interaction of nature and nurture in development: a parallel-distributed-processing perspective, in *Current Advances in Psychological Science: Ongoing Research* (Bertelson, P., Eelen, P. and D'Ydewalle, G., eds), pp. 57–88, Erlbaum

31 Hebb, D.O. (1949) *The Organization of Behavior*, John Wiley & Sons

32 Collingridge, G.L. and Bliss, T.V.P. (1987) NMDA receptors: their role in long-term potentiation *Trends Neurosci.* 10, 288–293

33 Oja, E. (1982) A simplified neuron model as a principal component analyzer *J. Math. Biol.* 15, 267–273

34 Landauer, T.K. and Dumais, S.T. (1997) A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge *Psychol. Rev*. 104, 211–240

35 Massaro, D.W. (1989) Testing between the TRACE model and the fuzzy logical model of speech perception *Cognit. Psychol*. 21, 398–421

36 Morton, J. (1969) The interaction of information in word recognition *Psychol. Rev*. 76, 165–178

37 Hinton, G.E. and Ghahramani, Z. (1997) Generative models for discovering sparse distributed representations *Philos. Trans. R. Soc. London Ser. B* 352, 1177–1190

38 Nowlan, S.J. (1990) Maximum likelihood competitive learning, in *Advances in Neural Information Processing Systems* (Vol. 2) (Touretzky, D.S., ed.), pp. 574–582, Morgan Kaufman

39 Jacobs, R.A. *et al.* (1991) Adaptive mixtures of local experts *Neural Comput.* 3, 79–87

40 Dayan, P. and Zemel, R.S. (1995) Competition and multiple cause models *Neural Comput.* 7, 565–579

41 Saund, E. (1995) A multiple-cause mixture model for unsupervised learning *Neural Comput.* 7, 51–71

42 Bienenstock, E.L., Cooper, L.N. and Munro, P.W. (1982) Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex *J. Neurosci*. 2, 32–48

43 Földiák, P. (1990) Forming sparse representations by local anti-Hebbian learning *Biol. Cybern.* 64, 165–170

44 Majani, E., Erlarson, R. and Abu-Mostafa, Y. (1989) The induction of multiscale temporal structure, in *Advances in Neural Information Processing Systems* (Vol. 1) (Touretzky, D.S., ed.), pp. 634–642, Morgan Kaufmann

45 Fukai, T. and Tanaka, S. (1997) A simple neural network exhibiting selective activation of neuronal ensembles: from winner-take-all to winners-share-all *Neural Comput.* 9, 77–97

46 Hinton, G.E. (1989) Connectionist learning procedures *Artif. Intell.* 40, 185–234

47 Baldi, P. and Hornik, K. (1989) Neural networks and principal components analysis: learning from examples without local minima *Neural Netw.* 2, 53–58

48 Dayan, P. *et al.* (1995) The Helmholtz machine *Neural Comput.* 7, 889–904

49 Becker, S. (1996) Mutual information maximization: models of cortical self-organization *Network: Comput. Neural Syst.* 7, 7–31

50 de Sa, V.R. and Ballard, D.H. (1998) Category learning through multimodality sensing *Neural Comput.* 10, 1097–1118

51 Kay, J., Floreano, D. and Phillips, W. (1998) Contextually guided unsupervised learning using local multivariate binary processors *Neural Netw.* 11, 117–140

52 Hancock, P.J.B., Smith, L.S. and Phillips, W.A. (1991) A biologically supported error-correcting learning rule *Neural Comput.* 3, 201–212

53 Geman, S., Bienenstock, E.L. and Doursat, R. (1992) Neural networks and the bias/variance dilemma *Neural Comput.* 4, 1–58

54 Weigend, A.S., Rumelhart, D.E. and Huberman, B.A. (1991) Generalization by weight-elimination with application to forecasting, in *Advances in Neural Information Processing Systems* (Vol. 3) (Lippmann, R.P., Moody, J.E. and Touretzky, D.S., eds), pp. 875–882, Morgan Kaufman

55 Jeong, D-G. and Lee, S-Y. (1996) Merging back-propagation and Hebbian learning rules for robust classifications *Neural Netw.* 9, 1213–1222