Softky WR, Koch C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J Neurosci* 13:334–350.

Sparks DL, Lee C, Rohrer WH. (1990). Population coding of the direction, amplitude and velocity of saccadic eye movements by neurons in the superior colliculus. In: *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 805–811.

Sporns O, Tononi G, Edelman GM. (1991). Modeling perceptual grouping and figure-ground segregation by means of active reentrant connections. *Proc Natl Acad Sci USA* 88:129–133.

Tanaka K, Saito H, Fukada Y, Moriya M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J Neurophysiol* 66:170–189.

Thomson AM, West DC. (1993). Fluctuations in pyramid-pyramid excitatory postsynaptic potentials modified by presynaptic firing pattern and postsynaptic membrane potential using paired intracellular recordings in rat neocortex. *Neuroscience* 54:329–346.

Ungerleider LG, Mishkin M. (1982). Two cortical visual systems. In: *Analysis of visual behavior*. DJ Ingle, ed. MIT Press, Cambridge, 564–586.

von der Malsburg C. (1985). Nervous structures with dynamical links. *Ber Bunsenges Phys Chem* 89:703–710.

von der Malsburg C, Schneider W. (1986). A neural cocktail-party processor. *Biol Cybernet* 54:29–40.

von Noorden GK. (1990). *Binocular vision and ocular motility. Theory and management of strabismus*. CV Mosby, St. Louis.

Wurtz RH, Yamasaki DS, Duffy DJ, Roy JP. (1990). Functional specialization for visual motion processing in primate cerebral cortex. In: *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 717–727.

Young MP. (1992). Objective analysis of the topological organization of the primate cortical visual system. *Nature* 358:152–155.

Zeki SM. (1973). Colour coding in the rhesus monkey prestriate cortex. *Brain Res* 53:422–427.

Zeki S, Watson JDG, Lueck CJ, Friston KJ, Kennard C, Frackowiak RSJ. (1991). A direct demonstration of functional specialization in human visual cortex. *J Neurosci* 11:641–649.

7 The Binding Problem of Neural Networks

C. von der Malsburg

#37

The life of a person or animal is a succession of scenes: arrangements of animate and inanimate objects and their trajectories in time. The survival depends in large part on an animal's skill in exploiting the regularities in scenes to its advantage. No two scenes in a lifetime are alike in detail, and their regularity never has the form of a literal repetition of sensory patterns. Therefore, relevant structure cannot be formulated directly on a primary sensory level. Rather, it is necessary for the brain to formulate abstract patterns, or schemata, that relate indirectly to concrete sensory or motor patterns by complex computations. The great flexibility of even the humblest animal in dealing with its environment shows that the relevant regularities are general. The brain's architecture must capture the style of these regularities, must provide a framework for making the right distinctions and identifications, and must provide for the organizational mechanisms to create appropriate representations and action patterns as well as for learning.

COGNITIVE ARCHITECTURE

The set of basic structures and mechanisms that enable the brain to extract and process the regularities behind sensory and motor patterns may be called its cognitive architecture. Four basic issues span much of the issue (figure 7.1):

1. The brain is a physical system. Its states can be interpreted as representations of scenes. What are the relevant physical quantities and how are they to be interpreted? *What is the data structure of short-term memory?*

2. Under the influence of sensory information and of stored information (i.e., the data referred to under issue 3), the contents of short-term memory are organized. *What is the mechanism of short-term memory organization?*

3. Activity in the brain leaves behind physical traces that can be interpreted as representing knowledge, experience, patterns, rules, procedures, skills, and memories. What is the nature of those traces and how are they to be interpreted? *What is the data structure of long-term memory?*

4. Depending on the state, that is, the contents of short-term memory, the contents of long-term memory are modified to store information that may be

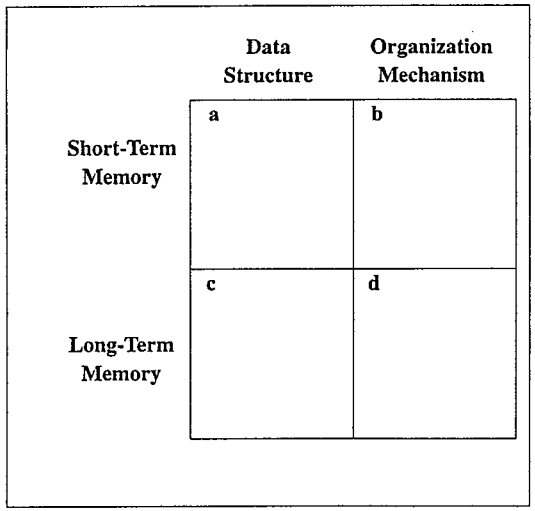

Figure 7.1 Cognitive architecture. Two fundamental questions must be addressed to understand the relationship between brain and mind: how are physical states of the brain to be interpreted as representing mental states (in computer terms, what is the brain's data structure)? and what are the mechanisms by which brain states are organized? Both questions require answers on at least two time scales: that of short-term memory, representing the present scene, and that of long-term memory, representing our knowledge and experience. These questions can be conveniently arranged as a 2 × 2 matrix. The letters a through d in the four boxes refer to the issues discussed in the text.

useful at a later time. What is the nature of this process? *What is the mechanism of long-term memory organization?*

These questions asked about the computer would not make much sense. Each particular program has its own data structures and algorithms that have to be interpreted one by one. There are general answers to those general questions, but they refer to a trivial level and are not very interesting. However, we have reason to believe that answers concerning the brain will be much more interesting.

The brain's architecture is the result of phylogenetic development. Phylogeny could not afford to pay attention to very specific situations, at least not for animals that are able to adapt flexibly to new environments. Consequently, it had to formulate structure on a principled, architectonic level, specifying application-oriented structures only on a general level and letting the individual adapt its nervous system to its particular needs.

CLASSIC NEURAL NETWORKS AS COGNITIVE ARCHITECTURE

The baseline from which to start is the conceptual framework currently discussed under the name of neural networks. This framework is the fruit of decades of conceptual development, and its main points are well grounded in neuroscience. Here is how it deals with the four architectural issues:

1. The physical variables relevant to short-term memory are neural firing rates. Each neuron can be interpreted as an elementary symbol. The meaning of that symbol can be investigated by recording from the neuron in the active brain and finding the precise context that activates it. Typical symbolic meanings of neurons are a blue light bar moving with orientation o over position x of retina, or muscle y twitched, and grandmother. The symbol is alive and part of the actual perceived scene when the corresponding neuron is active.

2. Activity states of neurons are organized with the help of excitatory and inhibitory signals exchanged between neurons and received from sensory cells. The system is regulated such as to converge toward stationary states. These last a fraction of a second, a time scale tuned to the typical progression of events in real life situations.

3. The physical quantities constituting long-term memory are synaptic strengths. If the customary definition is stretched a bit, one can interpret a synapse as an elementary rule: the excitatory connection from neuron j to neuron i, reading, when j is on, the probability that neuron i should also be on is to be raised by an amount related to the strength T_{ij} of the synapse between them.

4. Long-term memory is reorganized with the help of mechanisms of synaptic plasticity. One of these mechanisms is Hebbian plasticity, according to which the (excitatory) connection between two neurons is strengthened when both neurons are active at the same time. The effect of this rule is to increase the likelihood of those activity states that occurred in the past. Other plasticity rules implement supervised learning and use external information on the desirability of the present activity state, or partial information on which state should have been activated.

This set of concepts dominates current thinking about the function of the brain and has much to recommend it. It structures our thinking about the brain and raises relevant issues. It is in line with much of the experimental evidence about the nervous system and suggests further experiments in a fruitful way. Most important, it constitutes a vision that encourages us to formulate and attack the issue of cognitive architecture.

However, problems exist with classic neural networks as a candidate for cognitive architecture (Fodor and Pylyshyn, 1988). Although the formation of specific functional structures by learning from examples has been demonstrated, these demonstrations are restricted to small problem spaces and are successful only if the input patterns are encoded already in a way that is adapted to the problem at hand. (Neural networks have inherited these constraints from the more general framework of statistical estimation.) Thus, the technology of forming functional structures is still too much a matter of construction rather than self-organization. With their limitations, neural

network structures can serve only as small subsystems in a larger system, the construction of which is a matter of construction and not of autonomous learning.

Another closely related difficulty with classic neural networks (again, taken as a technological tool) is their lack of power to generalize. Thus, the greater goals of representing natural scenes or of organizing the behavior of a complex organism seem totally out of reach of neural network-based theorizing. The dynamic link architecture is an attempt to solve a fundamental problem with neural networks, binding, and may thus help to overcome some of their difficulties.

THE BINDING ISSUE

Imagine a neural network for the inspection of a visual scene as mediated by its image on the retina. The network is internally structured such that it can derive four propositions and represent them by output neurons. Two of them recognize objects: a triangle (neuron *triangle*) or a square (*square*), both generalizing position. The other two indicate the position of objects: in the upper half (*top*) or the lower half (*bottom*) of the retina, both generalizing the nature of the object. When shown single objects, the network responds adequately, with (*triangle, top*) or (*square, bottom*). A problem arises, however, when two objects are presented simultaneously. If the output reads (*triangle, square, top, bottom*) it is not clear whether the triangle or the square is in the upper position. This is the binding problem: the neural data structure does not provide for a means of binding the proposition *top* to the proposition *triangle*, or *bottom* to *square*, if that is the correct description. In a typographical system this could easily be done by rearranging symbols and adding brackets: [(*triangle, top*), (*square, bottom*)]. The problem with neural networks is that they provide neither for the equivalent of brackets nor for the rearrangement of symbols.

This example, due to Frank Rosenblatt, uncovers a fundamental problem with the classic neural network version of a cognitive architecture. The problem refers to issue 1, the data structure of short-term memory. Neural networks have no flexible means of constructing higher-level symbols by combining more elementary symbols [(in the example, the composite symbol (*triangle, top*) out of *triangle* and *top*)]. The difficulty is that simply coactivating the elementary symbols leads to binding ambiguity when more than one composite symbol is to be expressed. This weakness can have grave consequences. If it were vital for the organism to trigger some action in response to a triangle, but only if it was in an upper position, the scene representation given above would not be sufficient. The reaction would have to be tied to the coincidence of activity in cells *triangle* and *top*, which, however, would also occur if the triangle was at the bottom and a square was at the top. The animal therefore would respond to a so-called false conjunction.

It pays to analyze the origin of the binding problem in this example. The correspondence between object type and position is explicit on the retinal level. Its loss on the way to the output of the circuit is due to the generalization that is taking place within the circuit: the triangle and square cells perform generalization regarding position, the top and bottom cells perform generalization regarding object type. (In this sense, the circuit has its own "what" and "where" systems, in analogy to the response types discussed for the temporal and parietal pathways of primate cortex.)

Remaining in the classic neural network framework for a moment, the hole can be stopped by introducing combination-coding cells, a neuron that stands for the combination (*triangle, top*), for instance. This solution, however, is problematic on more than one account. Our nervous system cannot afford to contain combination-coding neurons to represent all possible bindings, combinatorics quickly leading to astronomical numbers. It also cannot be imagined that evolution has found a way of endowing our brain at birth with a complement of neurons that is affordable in size and yet covers all combinations that can ever play a role in our life. What remains is the potential of creating new combination-coding neurons by learning whenever they turn out to be important.

This is the route taken by most current neural network models. The idea runs like this. A feature combination that is to be represented by a neuron first has to be recognized as being important, and for that it has to be picked from all the other combinations that are present in scenes actually occurring. This is achieved with the help of scene statistics: if the combination (*top, triangle*) occurs more often than other combinations, then it must be important and is to be represented by a new neuron.

This approach runs into a serious problem. The number of possible feature combinations in a scene is very large, growing exponentially with the number of features in a sensory modality. To estimate frequencies of occurrence with statistical significance, the necessary number of scenes would be much larger than the animal experiences in its lifetime. For this quantitative reason the statistical estimation approach to the learning of combination-coding neurons works only in very small model systems. This problem creates a barrier to the scaling up of model systems that has to be broken before the dream of systems that learn from a natural environment can become real and before we can claim to understand the brain.

This learning time problem for the formation of combination-coding cells is due to the absence of a flexible, dynamic binding mechanism in classic neural networks. Such a mechanism could obviate the need for most of these cells, and it could also be extremely useful for identifying important combinations to be represented by neurons should the need arise. In the example above, the relevant binding information is still explicitly present at the retinal level—it only has to be handed down and represented at the output despite the generalizations taking place in the circuit. If that could be achieved somehow, learning could be speeded up tremendously because the significant

bindings would be made to stand out in individual scenes. In short, classic neural networks present us with the paradoxical situation that binding is possible with combination-coding cells, but combination-coding cells are not easily available without binding.

Conventional symbol systems all have flexible means of expressing binding and hierarchical composition of higher-level symbols from more elementary symbols. For that they mostly employ spatial or temporal arrangements of subsymbols. However, these solutions to the binding problem cannot serve as models here. The brain poses the added difficulty that its data structure should not only express binding but must also be structured by general mechanisms of self-organization.

Binding addresses very deep issues concerning the brain and the mind. When the physical structure of the brain is examined, one finds molecules, cells, and connections. When we examine the mind by introspection or with the help of psychophysics, we encounter entities that integrate large amounts of detail into global patterns, imaginations, and decisions. It is very likely that these two faces of the coin can be united in a view according to which the detail is identified with the activity of individual neurons, and the global aspects are identified with coherent dynamic patterns on the array of all neurons. Realization of this view requires a vision of how the collective dynamics of detail-encoding individual neurons can be integrated into higher-level mental entities. The classic neural network architecture tries instead to fold the mental entities back onto the single cell level.

DYNAMIC LINK ARCHITECTURE

The cells of the central nervous system are extremely complex. Classic neural networks are based on a simple caricature of this complexity. Perhaps this caricature leaves out essential traits, already known or yet to be discovered. They dynamic link architecture (von der Malsburg, 1981) modifies neural networks in five essential ways:

1. Rapid neural signal fluctuations are considered an essential aspect of neural activity, not as inessential noise.

2. Dynamic links are introduced as a new set of variables in the form of rapidly modifiable synaptic weights.

3. Synaptic and neural activity processes are interpreted in a novel way as data structures for short-term memory, issue 1 of the cognitive architecture.

4. A more complex organization mechanism for short-term memory is introduced (2).

5. The process of long-term synaptic plasticity (4) takes on a somewhat more refined form.

Several mathematical formulations have been attempted for these ideas (von der Malsburg, 1988; Bienenstock and von der Malsburg, 1987; Konen, Maurer, and von der Malsburg, 1994), but it would be premature to give preference to any one of them. Consequently, I will define the architecture here as a conceptual rather than as a mathematical structure.

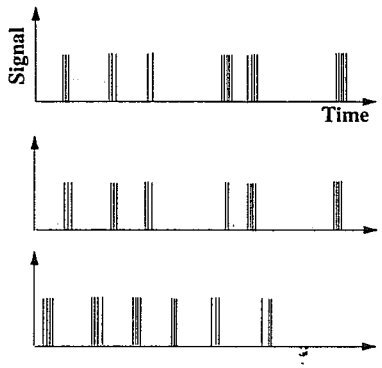

Figure 7.2 Temporal binding. Neurons can dynamically express grouping into composite structures by synchronizing their activity. In the figure, the middle neuron is bound to the upper neuron and not to the lower. If the middle neuron stood for a position and the other two for objects, the position would in the present situation apply to the upper object, not the lower. At another time, the lower neuron's object might be in that same position, expressed by spike synchrony between the middle and bottom neurons.

Correlations

Neurons in the brain are taken to be elementary symbols, just as in the neural network architecture. The signals of neurons are evaluated under two aspects. One of them is the rate or mean frequency of firing, and is interpreted as the intensity with which the elementary symbol is presently active. The other aspect refers to fine temporal signal structure, which is evaluated in terms of correlations among sets of cells. If, during a time interval, the signals on a set of neurons are found to be significantly correlated, the set is interpreted as being bound during that interval (figure 7.2).

There is experimental evidence for the existence in the nervous system of temporal signal structure of appropriate nature to encode binding by temporal synchrony. It has even been observed that signal correlations between cells occur and disappear when the features to which the cells respond are integrated into one figure or are part of distinct figures (e.g., two vertical light bars moving horizontally as one long bar or moving independent of each other). For a review of some of the experimental evidence see Engel et al (1992) and Singer, this volume.

The simplest case involves just two neurons and their binary correlation. It is to be borne in mind, however, that correlations involving just two neurons cannot play a central role in the brain. Binary correlations are too easily

drowned in the noise of accidental coincidences. Therefore, more important correlations will be those involving larger numbers of neurons. A few events, each involving simultaneous spikes on fifty neurons, can easily be recognized as being statistically significant. A variety of possible temporal signal structures could serve as a basis for correlation patterns, with irregular spike trains at one end of a spectrum and regular oscillatory signals (Engel et al, 1992) at the other. In the latter case, correlations between two signals express themselves as agreement in frequency and phase.

This interpretation is to be applied on a hierarchy of several time scales. An event at a coarser time scale is composed of a rapid succession of short events. The hierarchy is limited below by the precision with which relative timing of signals can be reproduced in the central nervous system. This may be somewhat above one millisecond.

The hierarchy of time scales corresponds to a hierarchy of complexity. Speaking about the visual system, on the finest scale neurons may be bound together that are activated from the same point on a retina and that refer to different submodalities such as shape, color, motion, or stereo depth. On a higher level of the binding hierarchy, neurons in the same neighborhood within a figure are correlated with somewhat less temporal precision. Binding on this level has been invoked to bind together the elements of shape primitives, or geons (Hummel and Biederman, 1992). On a next higher level, cells anywhere within the same figure are synchronized. If, for instance, two colored letters are seen side by side, the spatial resolution of color-sensitive neurons may not be good enough to distinguish between the positions of the two letters. Without a binding mechanism, this would lead to ambiguity as to which color goes with which letter. The corresponding "conjunction error" is actually observed when subjects are not given sufficient viewing time (Treisman, 1985). With temporal binding, the signals of all those neurons that are activated from the same retinal point are correlated in time, disambiguating the situation. This would also solve the binding problem in the triangle-and-square example.

On a still higher level of the hierarchy, entire objects are bound to form the representation of a whole scene or sentence or complex argument. It has been argued that the phase relations of oscillatory firing can be employed to implement reasoning in neural networks (Shastri and Ajjanagadde, 1993).

The lower end of the binding hierarchy is not accessible to our own introspection. At scales larger than, say, 0.5 second our brain is able to observe and remember the actual sequence of events. They are experienced as the coherent chunks of structure that appear in flashes of attention. On coarser time scales, events are to a large part staged by eye and body movements and by real external events. On these scales the phenomenon of binding by simultaneous activity has always been taken for granted. It is also implicit in the usual way of driving neural networks as a sequence of stimulus-response events: simultaneously active input units are thereby bound as part of the same stimulus. The dynamic link architecture simply continues the hierarchy to finer temporal scales and thereby to events that are created by the network itself rather than the input.

On the scale of a few milliseconds, signal correlations may actually not be restricted to precise synchrony, and the spikes that form an event may be slightly scattered in time. It is necessary, however, that the actual sequence be reproducible and that the circuitry be in place to distinguish the temporal patterns.

One could go to the extreme and imagine a system that started out entirely composed of low-level sensory and motor neurons and that built up all higher-level entities as connectivity and correlation patterns. Objects would then be represented in a natural way as arrays of their initially given sensory elements, the same way that the tea mug in front of me is constituted entirely by its atoms and does not contain any high-level units to correspond to its subpatterns, its parts, or its entirety. This view in its extreme form is certainly not realized in the brain. The bandwidth of neural signals in our nervous system is too low to permit the build-up of very deep correlation hierarchies. For this reason, intermediate levels in symbol hierarchies have to be represented to a large extent by new units that are to be recruited along the way, temporal correlations stepping in to represent those combinations that are not (yet) represented by cells.

It has been claimed repeatedly that the observation of combination-coding neurons in a brain is evidence for a solution to the binding problem without temporal coding. This argument misses the point. The binding problem arises only when the combination-coding neurons in the brain cannot disambiguate the situation.

Neurophysiologists have traditionally considered random aspects of cellular responses to test stimuli as a nuisance, and they averaged them out by adding responses over many identical stimulations. Also, modeling studies of neural activity in the brain tended to ignore fine temporal signal structure (and, *a fortiori*, their correlations) with the following argument. The number of synapses converging onto a neuron is very large. Under the assumption that the signals carried by them are statistically independent, their temporal structure will average out, resulting in a temporally smooth summed input signal to the neuron. Correspondingly, the output of the neuron will vary smoothly, and this would be true for all neurons. Any temporal signal structure would thus quickly disappear. According to this argument it seems to be a self-consistent view to expect neural signals to vary smoothly in time. However, this view does not stand up to fact. As any recording shows, cortical neural signals do have a very pronounced stochastic structure, and nervous tissue is evidently designed to create and preserve it. The assumption of statistical independence of signals therefore must be faulty, and strong signal correlations must be present in the brain.

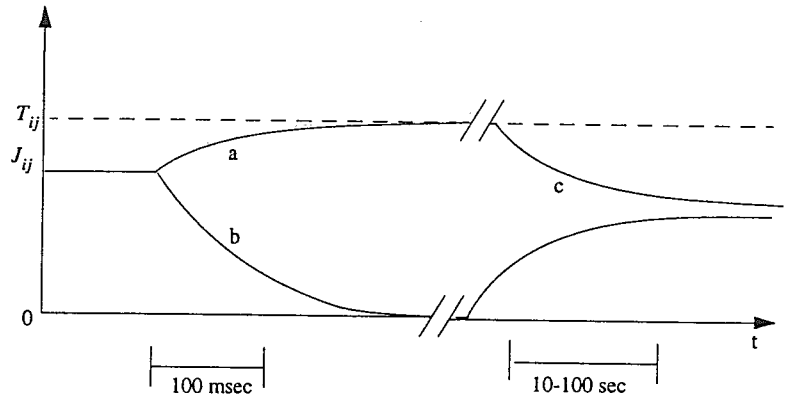

Figure 7.3 Dynamic links. The synaptic weight J_{ij} between neurons j and i can vary on a fast time scale. (a) If the neurons i and j fire synchronously, the dynamic weight J_{ij} is rapidly increased from a resting value to the maximum value set by the permanent synaptic strength. (b) If both neurons are active but fire asynchronously, the synaptic weight is rapidly decreased to zero. This dynamic switching can take place within small fractions of a second. (c) When there is no more activity in one or both of the neurons, the dynamic weight slowly falls back to the resting value, with the time constant of short-term memory.

Rapid Synaptic Modification

The temporal bandwidth of neural signals is severely limited, and it is not possible within short periods of time to express complicated multilevel binding structures in terms of signal correlations. If it is necessary to do so, longer periods of time have to be invested to process the many disambiguations sequentially. If it has been observed that two neurons are actually not bound in the present scene, that observation can conveniently be stored in short-term memory by temporarily switching off any connections that may exist between them and to common target cells.

This observation leads to the hypothesis of rapid synaptic modification (figure 7.3). Synapses are characterized by two quantities, T and J, where T is the conventional permanent synaptic weight, which may be slowly modified by mechanisms of plasticity (see below). The momentarily effective strength of a synapse is J. This quantity can vary on a rapid time scale between zero and a maximum determined by T. When J is zero, the synapse does not transmit signals. At rest, J takes on an intermediate value between zero and its maximum. When the two connected cells become active in a correlated fashion, J is regulated upward, to its maximum if the correlation is sufficiently strong. When the two cells are both active but in an uncorrelated fashion, J is regulated downward, to zero in the extreme case. Forceful activity events can modify synapses in a small fraction of a second, perhaps a few milliseconds. When there is no further activity in the two cells connected by a synapse, J slowly returns to its resting value, with the time constant of short-term memory.

Although experimental evidence exists for rapid synaptic modification (Zucker, 1989), the precise mode of control postulated in dynamic link architecture remains to be demonstrated in experiments. The effective synaptic weights of whole fields of synapses can be reset to their resting values by central command signals.

The two central hypotheses of dynamic link architecture, binding by signal correlations and short-term synaptic modification, amount to a massive increase in the number of dynamic variables, to a much richer structure and to a profound reinterpretation of neural dynamics as part of the cognitive architecture. Signal correlations are the glue by which high-level symbols can be formed from lower-level ones to suit the needs of individual scenes and the constellation of objects and patterns in them. If, as done above, synaptic connections are interpreted as elementary rules, then by dynamically switching synapses, the system can decide which of the rules are actually applicable in a given situation, switching off those that are not consistent with the situation and with each other. To manage all this rich structure, more sophisticated mechanisms of organization have to be considered (to deal with issues 2 and 4 of the cognitive architecture).

State Organization

In the classic neural network architecture, the only variables on the psychological time scale are neural signals. These are organized by the exchange of excitation and inhibition, regulated to create attractor states that are quasi-stable over fractions of a second. In the dynamic link architecture, signal dynamics are modified to create signals that fluctuate on several time scales. This raises a number of issues, especially questions of how functionally appropriate correlations are created and how correlations are used by the brain in a meaningful way.

Some of the signal correlations are already implicit in the sensory signals and express the causal relations in the external world. More correlations are created, however, in the neural circuits themselves: the existence of excitatory connections between two neurons raises the probability for one of them to fire in correlation with the other. The essence of this is that the signal correlations come to express the patterns of connections in the circuit, that is, the underlying causal structure of the nervous system itself, just as afferent signal correlations express the causal structure in the external world. The "fast enabling links" (Hummel and Biederman, 1992), for instance, are connections within the visual system that encode those relations of features that are indicative of binding within a geon.

How are signal correlations evaluated in the brain? According to what we know about the structure of the nervous system, signal correlations play a very important role in neural dynamics. The excitatory effect of afferent

spikes can pile up and fire a cell only if they coincide with a precision comparable to the membrane time constant. Thus, neurons are coincidence detectors and are superbly sensitive to signal correlations. Now, if coincidence-detecting neurons were the only means of evaluating correlation, one would not be much better off than classic neural networks, and the brain would still need a combination coding cell for each binding pattern. Fortunately, that is not so. A correlation pattern can selectively activate receiving circuits of appropriate structure. In the simplest case this is possible if the circuit that receives the pattern is isomorphic to the circuit that created it, in the sense that subsets of coupled cells in one circuit are connected to subsets of coupled cells in the other. This principle was extensively exploited for the invariant recognition of visual patterns (Konen, Maurer, and von der Malsburg, 1994).

Whereas in classic neural networks neurons do not pay attention to fine signal structure, in the dynamic link architecture the incoming signals have to be correlated in time. This leads to a much more differentiated dynamics, and a given set of neurons can now support a large number of activity patterns that differ only in their fine temporal structure. This principle is also implicit in Abeles' (1991) synfire chains. These are chains of sets ("pools") of neurons, each one connected nearly all-to-all with the next pool. A synfire chain can be traversed by an activity process in which all the cells in a pool fire simultaneously and thus succeed in firing the cells in the next pool simultaneously. Each neuron can participate in many synfire chains and even several times in the same chain. Synfire chains are proposed as the basic building blocks of a compositional cognitive system (Bienenstock, 1994).

Signal correlations are also evaluated by synapses, by modifying their dynamic weight. The interaction between signal correlations and synaptic dynamics has the form of a positive feedback loop. A strong synapse helps to create a correlation in the two cells connected, and the correlation strengthens that synapse. This feedback loop is the basis for a system of rapid network self-organization. A given network creates a signal process that is characterized by correlations shaped by the connectivity structure in the network. These signal correlations act back on the network and modify its structure by rapid synaptic modification. This leads to a run-away situation that comes to a halt when a network is reached in which the signal structure and the connectivity structure are consistent with each other. These specific structures are called connectivity patterns. Presumably (von der Malsburg, 1981), connectivity patterns are sparse, that is, have relatively few active connections per node; and the active connections are maximally cooperative, that is, for each pair of neurons in the pattern, different direct and indirect connections help each other to link those neurons.

Scene Representation

The reality we perceive is first and foremost the reality of the states of our mind. It is for us an issue of central importance to find out how the physical states of our brain can come to represent that rich world. This is exactly the architectural issue 1: how can activity states or processes in our brain act as physical symbols to represent the mental objects that we see and experience? I will argue that the dynamic link architecture is a major step toward solving that issue due to its property of compositionality (see Bienenstock and Geman, this volume), which distinguishes it from classic neural network theory.

Our inner experience can be described as a continuous sequence of scenes. Voluminous tomes have been written to describe scenes from an introspective point of view (an extreme example is the work of E. Husserl). Briefly, a scene is a short, real or imagined sequence of events that are simultaneously and explicitly accessible, and usually contain a description of ourselves as part of the scene. Scenes can be decomposed into separate objects, their trajectories, and their relationships. Mental scene descriptions are constructed (issue 2) from sensory data and from stored knowledge (issue 3). We must at the very least understand how scenes can be composed from simpler entities and descriptors, how they can be decomposed, and how new patterns can be referred back to known, old patterns.

Presumably, our mental system is composed of relatively fixed elementary patterns that can be dynamically linked to form more complex descriptors. To take an example, my tea mug is described in terms of shape aspects (elongated, upright form, cylindrical body, handle), shape elements (curvature of surface, flared rim, bend of the handle), surface markings (flower decor, shiny reflections), its material, its position relative to me and the table as reference surface, its function as a container of liquid and source of vapor, its role in the potential action of grasping and drinking, and, if I care to elicit them, many more aspects.

The dynamic link architecture affords the infrastructure required for scene representation. Complex (sensory) patterns can be segmented in a meaningful way into subpatterns that correspond to separate objects or functional components of the scene. This is done by temporally correlating signals within a segment and decorrelating signals between elements in different segments. During the process of segmentation, experience on likely groupings of elements, stored in the form of permanent connections, can be brought to bear (Hummel and Biederman, 1992). Segmentation in the context of the dynamic link architecture has been described by a number of authors (von der Malsburg and Buhmann, 1992).

Whereas the formation of segments is possible with the help of very simple signal patterns in the form of blocks uniting all elements within a segment in an undifferentiated correlation, the description of the internal structure of an object requires detailed correlation patterns that amount to intricate arrays of pointers that attach descriptors to their referents. This attachment is achieved with the help of binary signal correlations between descriptors and referents.

A paradigm of object representation and the attachment of descriptors was studied in the context of visual object recognition (von der Malsburg, 1988;

Konen, Maurer, and von der Malsburg, 1994). The two-dimensional visual aspect of an object is constructed by linking neural elements in the form of a two-dimensional array. Individual elements represent wavelet components (roughly, oriented edges on various levels of resolution) of the gray-level distribution on the retina. Such elements are found as receptive fields of single neurons in primary visual cortex. A collection of such visual aspects is stored as models in some part of the brain ("model domain") as networks of permanent connections. An element can take part in many such networks. In a concrete situation all the links belonging to one aspect are dynamically activated and all others are deactivated. This is possible without creating confusion (Bienenstock and von der Malsburg, 1987). When a new aspect of an object appears in the primary visual cortex (and is naturally represented as a two-dimensionally linked array of wavelet elements), it can be recognized with the help of a dynamic process in which an appropriate aspect model is activated and its parts are linked to the new visual aspect, attaching corresponding elements to each other (Konen, Maurer, and von der Malsburg, 1994). An appropriate model is one that is loosely isomorphic to the new aspect in containing the same element types in the same two-dimensional arrangement. Object recognition that is invariant to position, size, and orientation and that is robust with respect to lighting, background, partial occlusion, deformation, and rotation in depth, and that can reliably distinguish between hundreds of objects (human faces) has been demonstrated in this style. Part of this work is described and reviewed by Lades et al (1993). In this way, a scene is interpreted and built up by the activation and interlinking of structural descriptions, attaching them to each other as suggested by structural relations of partial isomorphy. As this interlinking is represented with the help of temporal signal structure and correlations, these physical symbols for mental entities are processes rather than static structures.

It is often necessary for the brain to process temporally structured input patterns, especially in the auditory modality. These temporal patterns are in danger of colliding with the temporal processing required by the dynamic link architecture. This conflict is resolved with the help of peripheral circuits that detect temporal patterns in the sensory input and represent them with the help of slowly varying signals on specialized neurons, thus freeing the temporal domain on central levels from the interference by external patterns.

Learning

The amount of genetic information to structure the brain is limited, and especially for humans, genes cannot address the living environment of the individual in a very specific way. The brain therefore has to build up autonomously or absorb much of the required structure. The vision motivating the field of neural computing is to understand and imitate this ability. To date this goal is distant, and all neural model systems still rely heavily on very specific initial connectivity structure. The dynamic link architecture may bring the vision within reach, by solving a fundamental difficulty of neural learning.

So far I have discussed the organization (issue 2) of the active state of the system (issue 1), making use of the constraints implicit in the permanent connectivity parameters T, which constitute the long-term memory (issue 3). An important modifaction is also introduced with respect to modification of the long-term memory (issue 4). The mechanisms for modifying synaptic connections in classic neural networks architecture, Hebbian plasticity, reinforcement learning, and supervised learning, are all plagued by the difficulty that the significant and meaningful connections that have to be modified are easily drowned among the many meaningless connections that correspond to accidental patterns active in a given state. As discussed above, this leads to bad scaling behavior of neural networks. Efficient learning requires mechanisms to single out those connections in a situation that are essential and significant, and distinguish them from those arising from accidental coexistence.

The dynamic link architecture offers a potent mechanism to do that. If, in a given state, two cells are active, and between them is neither a direct nor a short indirect connection, then the activity on these cells will indicate this fact by not being correlated. A plasticity mechanism that is sensitive to the fine signal correlations, called refined plasticity, easily picks up this fact and keeps them apart, neither creating a direct connection between them nor helping to create a hidden unit to represent a pattern that includes them. The power of this effect was demonstrated by Konen and von der Malsburg (1993).

Expressed in more general terms, the activity and connectivity patterns created in the dynamic link architecture constitute differentiated structure in a given state that goes way beyond merely expressing which cells are active in the state (as does the classic neural network architecture), making evident the way in which these cells are connected according to the knowledge already implicit in the network. This information is to a large extent ignored in the classic architecture. In the dynamic link architecture it is used to keep those connections from growing for which there is not at least indirect evidence of functional significance in the form of indirect connections linking the same end points.

Another way of seeing how the dynamic link architecture improves the learning situation is this. A given state of the neural network is broken down into a quick succession of microstates, each of which activates only a small subset of the cells in the state. Plasticity is restricted to act only within microstates. In this way the plasticity mechanism is saved from the need to scale to states with many active cells. The burden falls on the dynamic mechanism to break the network's state into microstates in a significant way.

REFERENCES

Abeles M. (1991). *Corticonics: Neuronal circuits of the cerebral cortex.* Cambridge University Press, Cambridge.

Bienenstock E. (1994). *A model of neocortex.* Technical report. Division of Applied Mathematics, Brown University, Providence, RI.

Bienenstock E, von der Malsburg C. (1987). A neural network for invariant pattern recognition. *Europhys Lett* 4:121–126.

Engel AK, König P, Kreiter AK, Schillen TB, Singer W. (1992). Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends Neurosci* 15:218–226.

Fodor J, Pylyshyn ZW. (1988). Connectionism and cognitive archtecture: A critical analysis. *Cognition* 28:3–71.

Hummel JE, Biederman I. (1992). Dynamic binding in a neural network for shape recognition. *Psychol Rev* 99:480–517.

Konen W, Maurer T, von der Malsburg C. (1994). A fast dynamic link matching algorithm for invariant pattern recognition. *Neural Networks* 7:1019–1030.

Konen W, von der Malsburg C. (1993). Learning to generalize from single examples in the dynamic link architecture. *Neural Computation* 5:719–735.

Lades M, Vorbrüggen JC, Buhmann J, Lange J, von der Malsburg C, Würtz RP, Konen W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans Comput* 42:300–311.

Shastri L, Ajjanagadde V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings. *Behav Brain Sci* 16:417–494.

Treisman A. (1985). Preattentive processing in vision. *Comput Vision Graphics Image Processing* 31:156–177.

von der Malsburg C. (1981). *The correlation theory of brain function.* Internal report 81-2. Max-Planck-Institut für Biophysikalische Chemie, Göttingen, Germany. Reprinted (1994). In: *Models of neural networks.* K Schulten, HJ van Hemmen, eds. Springer-Verlag, Berlin.

von der Malsburg C. (1988). Pattern recognition by labeled graph matching. *Neural Networks* 1:141–148.

von der Malsburg C, Buhmann J. (1992). Sensory segmentation with coupled neural oscillators. *Biol Cybernet* 67:233–242.

Zucker RS. (1989). Short-term synaptic plasticity. *Annu Rev Neurosci* 12:13–31.

8 Recognition and Representation of Visual Objects in Primates: Psychophysics and Physiology

N. K. Logothetis and D. L. Sheinberg

THE PROBLEM OF RECOGNITION

The ability to recognize objects is a remarkable accomplishment of biological systems. Familiar objects can be readily recognized based on their shape, color, or texture. Even when partially occluded, an object's identity can be deduced based on contextual information. Furthermore, visually similar members of object classes can become easily disciminable by repeated exposure, as when a geologist learns to recognize rock formations or an ornithologist learns to discriminate species of birds.

Reliable artificial recognition systems have proved to be surprisingly difficult to achieve. A major obstacle in this endeavor is that we know very little about what actually constitutes an object. Components of objects are not clearly labeled as belonging to one object or another. Indeed, there is nothing special about individual objects in the way they are presented to the visual system. The shape of an object, or its characteristic regions, are almost never visually primitive constructions, determined by a predictable combination of primary cues. Any given two-dimensional image can be parsed into an arbitrary set of objects, each of which can be decomposed recursively into smaller objects. Moreover, what we consider to be an object depends on the visual input, yet it is also determined by the task at hand.

The neural representation of objects is a mystery even when considering simple geometrical objects, such as a cube, a cone, or a cylinder, seen in isolation. A key question concerning the perception of three-dimensional objects is the spatial reference frame used by the brain to represent them. The rapidity of the recognition process could be explained by the visual system's ability to transform stored models of three-dimensional familiar objects quickly, or by its ability to specify the relationship among viewpoint-invariant features or volumetric primitives that can be used to accomplish a structural description of an image. Alternatively, viewpoint-invariant recognition could be realized by a system endowed with the ability to perform an interpolation between a set of stored two-dimensional templates created for each experienced viewpoint.