Psychonomic Bulletin & Review
1997, 4 (3), 299–309

Nature, nurture, and the development of functional specializations: A computational approach

ROBERT A. JACOBS
University of Rochester, Rochester, New York

The roles assigned to nature and nurture in the acquisition of functional specializations have been modified in recent years due to increasing evidence that experience-dependent processes are more influential in determining a brain region's functional properties than was previously supposed. Consequently, one may study the developmental principles that play a role in the acquisition of functional specializations. This article studies the hypothesis that a combination of structure–function correspondences plus the use of competition between modules leads to functional specializations. This principle has been instantiated in a family of neural network architectures referred to as "mixtures-of-experts" architectures. These architectures are sensitive to structure–function relationships in the sense that they often learn to allocate to each task a network whose structure is well matched to that task. The viewpoint advocated here represents a middle ground between nativist and constructivist views of modularity.

The concept of modularity is central to modern theories of the mind and brain. Indeed, the notion of modularity motivates significant portions of current research in the cognitive neurosciences, including research on perception, language, motor control, memory, and neural systems organization. These investigations often address at least two major theoretical and empirical issues. The first issue concerns the modularization of cognitive and behavioral faculties. Researchers seek to discover the extent to which different brain regions are specialized to perform different functions and the extent to which seemingly different behaviors have distinct underlying functional and neural processes, as well as to assess whether a given set of functional specializations is logical or efficient from an information processing viewpoint. The second issue concerns the acquisition of functional specializations. Researchers studying acquisition want to know whether the developmental processes that determine the functional properties of a brain region operate according to fixed genetic instructions or whether these processes are also experience sensitive.

The roles assigned to nature and nurture in the acquisition of functional specializations have been modified in recent years. A common view in psychology during the 1980s was that genetic factors played an exclusive role, or at least an overwhelmingly significant role, in determining the functional modules that characterize the human mind and in determining the specific brain regions that subserve each of the modules. For example, an emphasis on genetic determinism and the existence of stipulated modules can be found in Fodor's (1983) writings regarding so-called input systems (e.g., perceptual systems, motor systems, language systems). Fodor wrote,

> No facts now available contradict the claim that the neural mechanisms subserving input analysis develop according to specific, endogenously determined patterns under the impact of environmental releasers. This picture is, of course, quite compatible with the view that these mechanisms are instantiated in correspondingly specific, hardwired neural structures. It is also compatible with the suggestion that much of the information at the disposal of such systems is innately specified. (Fodor, 1983, p. 100)

Recently, however, this strong nativist view has been called into question due to increasing evidence that experience-dependent processes are more influential in determining a brain region's functional properties than was previously supposed.

Much of this evidence comes from studying cortical localizations of cognitive functions in human patients. For example, Ojemann, Ojemann, Lettich, and Berger (1989) found substantial individual variability in the exact cortical location of language function between patients and, thus, concluded that language cannot be reliably localized on the basis of anatomic criteria alone. Related evidence is provided by studies of hemidecorticate infantile hemiplegics (patients who had either the left or right hemisphere removed soon after birth due to cerebral injury). Dennis and Whitaker (1976) found that phonemic and semantic linguistic abilities are similarly developed in patients who have had either their left or right cerebral hemisphere removed, though syntactic competence is superior in patients with intact left hemispheres (see also Dennis & Kohn, 1975). Additional evidence has been obtained in studies comparing the behavior and event-

I thank I. Fine, S. Kosslyn, E. Newport, and J. Saffran for commenting on an earlier version of this manuscript. This work was supported by NIH Grant R29-MH54770. Correspondence should be addressed to R. A. Jacobs, Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627 (e-mail: robbie@bcs.rochester.edu).

related brain potentials (ERPs) of hearing adults and congenitally deaf adults during the performance of visual attentional tasks. Neville and her colleagues (summarized in Neville, 1995) found that ERPs to foveal stimuli were similar in congenitally deaf and hearing adults; however, ERPs over superior temporal cortical areas to peripheral stimuli were two to three times larger in deaf than in hearing subjects. Deaf adults also responded faster than hearing subjects in tasks requiring detection of movement in peripheral stimuli, though response times did not differ when foveal stimuli were used. On the basis of these results, Neville hypothesized that the portion of the visual system that mediates the processing of peripheral stimuli may, through a process of competitive interactions, take over brain regions in the congenitally deaf that would normally be auditory cortical fields either in primary sensory or multimodal cortical areas.

Further evidence of the experience-dependent nature of the acquisition of functional specializations comes from the study of developmental neurobiology. O'Leary (1989) argued that adult mammalian neocortex consists of numerous distinct areas, whereas developing neocortex lacks many of these area-specific distinctions. He reviewed evidence that this less differentiated structure undergoes considerable experience-dependent modification after neurogenesis that results in the emergence of well-defined neocortical areas. An example of experience-sensitive acquisition of functional properties was provided by Sur and his colleagues (e.g., Sur, Pallas, & Roe, 1990). These researchers induced retinal afferents to project to the medial geniculate nucleus (MGN), also referred to as "auditory thalamus." Consequently, visually responsive cells were recorded in MGN. MGN projects to primary auditory cortex and visually responsive cells were also found in this region. These cells tended to have large receptive fields, with roughly one third of the fields being orientation selective and a similar proportion being direction selective. Similar to the fields of simple or complex cells in normal visual cortex, the oriented receptive fields had either separate or coextensive *on* and *off* zones. In addition, many cells were driven binocularly. These results support the hypothesis that "primary sensory areas arise from regions of developing neocortex that are initially similar or to some extent pluripotent" (O'Leary, 1989, p. 401).

Nearly all of the research investigating the modular nature of the brain in general, and the acquisition of functional specializations in particular, is behavioral or neuroscientific in character. In contrast, my colleagues and I have been using computer simulations in order to study principles that might underlie the development of functional specializations. The framework that we have developed relies on two basic notions. The first notion is that there exist structure–function correspondences in the brain. Because different brain regions have different structural properties (e.g., different patterns of connectivity among their neurons), different regions are best at performing different types of functions. The second notion is that, analogous to Darwinian evolutionary processes, brain regions compete for the ability to perform a set of tasks. Regions become functionally specialized due to the competition; that is, different regions learn to perform different functions. Most importantly for our purposes, structure–function correspondences serve to bias the competition; each region tends to win the competition for those functions for which its structure makes it particularly well suited.

These two notions are not original to the computational framework reviewed in this article. The idea that competition leads to functional specializations has appeared in the literature in the form of the hypothesis that hemispheric specialization in humans is due to competition between neural subsystems. For example, Kosslyn (1987) proposed that subsystems of the brain compete to learn about inputs. If the output of a subsystem is used in subsequent computational processing, then the strengths of the neural connections in that subsystem are altered so that the subsystem produces the output faster and with less noise when the input recurs in the future. Subsystems whose outputs were not used in subsequent processing remain unchanged. Theories of brain lateralization also include the hypothesis that structure–function correspondences in the cerebral hemispheres may influence the lateralization of brain functions (Geschwind & Galaburda, 1987). For example, if the left and right hemispheres compete for the ability to process language, then anatomical differences between the two hemispheres may bias the competition so that the left hemisphere typically wins.

This article reviews a novel computational framework that implements a particular instantiation of these ideas. A contribution of this work is that it details, evaluates, and elaborates these ideas in a more explicit manner than has previously been possible. The computational implementation uses a family of neural network architectures referred to as "mixtures-of-experts" (ME) architectures, as well as a corresponding family of learning rules. These architectures consist of a number of loosely coupled modules. Adaptation in these systems is a combination of associative learning and competitive learning. Simulation results using these architectures show that it is possible to develop functionally specialized modules in an experience-dependent manner; that is, genetic stipulation is not required. However, genetic factors do play a role in that they influence the architectural structure of each module, thereby biasing, though not strictly determining, the development of each module's specific functional specialization. Because the functional specializations developed by the modules of an architecture are highly sensitive to the experiences of that architecture, and because genetic factors are moderately strong influences on the development of the specializations, the ME framework is compatible neither with a strong empiricist

view, nor with a strong nativist view, but rather with a view that emphasizes both environmental and genetic contributions to the behavior of a learner.

The article is organized as follows. First it overviews neural network models in general, and the ME family of architectures in particular. The next section reports the results of a number of simulation studies using a variety of ME architectures. My discussion of these studies does not emphasize modeling particular sets of experimental data (though the reader interested in this aspect of the work is encouraged to see the original articles). Rather, the emphasis is on showing the logical coherency, as well as some possible elaborations, of theories that attempt to account for the development of functional specializations via a combination of structure–function correspondences plus the use of competition between modules. The following section overviews some variants and extensions of the basic ME architecture, including a system in which some, but not all, of the modules participate in the competition, and also a hierarchical system. The final section relates the approach described here to a recent trend in developmental psychology toward theories that advocate a middle ground between nativist and constructivist views of modularity.

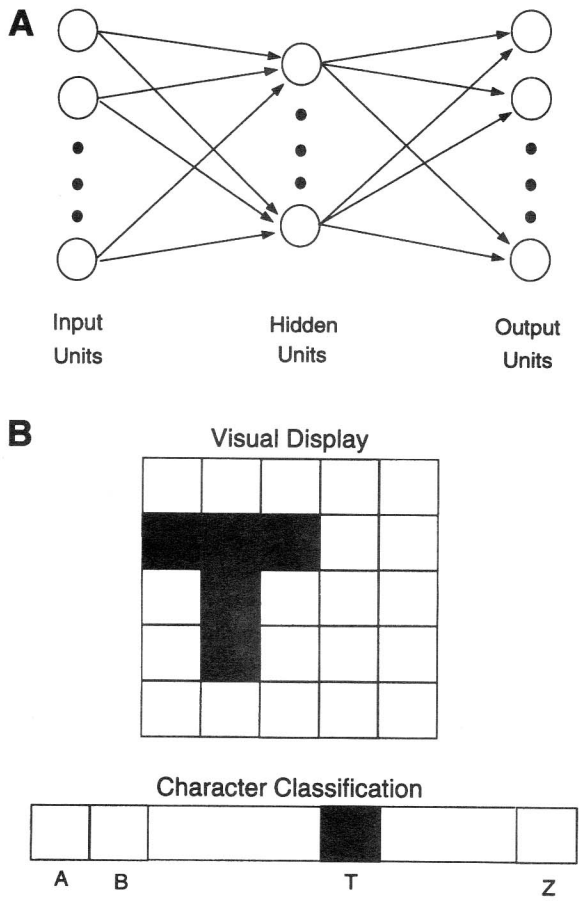

Figure 1. (A) A three-layered neural network. (B) A visual display of a character and the character classification.

ME ARCHITECTURES

Neural networks, also known as connectionist networks or parallel distributed processing models, have become over the past decade a popular tool for studying cognitive and neural processes (e.g., Quinlan, 1991; Rumelhart, McClelland, & the PDP Research Group, 1986). Each network consists of a number of simple processing elements, referred to as units, that are connected to each other to form a weblike structure. An example of a network is shown in Figure 1A. Each unit has associated with it an activation value that is typically a real number between 0 and 1. Associated with each connection between two units is a real-valued strength or weight. The activation of a unit is computed on the basis of the weighted sum of the activations of the units that project to it. Learning in neural networks involves adapting the connection strengths so that the network performs a desired associative task.

As an illustration, suppose that we would like a network to perform a character recognition task. At each time step, a letter is displayed on a panel that is 5 pixels wide and 5 pixels high (see Figure 1B). The network must decide which one of the 26 letters is currently displayed. Suppose that we use a network whose structure is similar to the one in Figure 1A. It includes 25 input units, 1 for each of the 25 pixels in the visual display. The activation of each input unit is set to 1 if its corresponding pixel is black, and set to 0 if its corresponding pixel is white. The input units send projections to a number of hidden units that, in turn, send projections to 26 output units. The target activation of an output unit is 1 if the unit corresponds to the visually displayed letter; otherwise it is 0. Most learning algorithms for neural networks work by adjusting the network's connection strengths until the output unit activations approximate their target values (e.g., Rumelhart, Hinton, & Williams, 1986). In our example, the network's connection strengths would be adjusted at each time step so that the output unit corresponding to the letter that is currently displayed has an activation near 1 and the activations of all other output units are near 0. An interesting feature of neural networks is that it is possible to examine their inner workings in order to determine how they have learned to perform a task. For the character recognition task, networks typically develop hidden units that act as "visual feature detectors." For example, a hidden unit may have a large activation when the display contains a line at a particular orientation; otherwise it has a small activation. Different units become tuned to lines at different orientations (Rueckl, Cave, & Kosslyn, 1989).

The most common neural network architecture is a single network in which a set of input units connect to a set of hidden units that connect to a set of output units. For our purposes, it is useful to note that this generic architecture is relatively unstructured, and its internal components are relatively undifferentiated. Researchers in the neural network community often implicitly advocate nonmodular processing systems, as evidenced by the frequent application of these generic networks to the

study of a wide variety of complex cognitive tasks. This is, at least in part, what critics mean when they accuse neural network researchers of adopting a tabula rasa approach to the study of mind (see Pinker & Prince, 1988). In contrast, my colleagues and I have advocated the use of highly structured modular architectures in which different networks learn to perform different tasks. An important property of modular architectures is that they can take advantage of task decompositions, meaning that a difficult task can be decomposed into a set of simpler subtasks and each network can learn to perform one of the subtasks. Modular architectures learn to perform tasks faster than single networks because task decompositions reduce the complexity of the function learned by each network of the architecture. Furthermore, suitably designed modular architectures also show better generalization, more interpretable representations, and more efficient use of hardware (Jacobs, Jordan, & Barto, 1991).

My colleagues and I have developed a modular architecture, referred to as the mixtures-of-experts (ME) architecture, that learns task decompositions in the sense that it uses different networks to learn input–output training patterns from different regions of the input space (i.e., the space of all possible inputs). There are two technical issues addressed by the ME architecture: (1) detecting that different training patterns belong to different tasks and (2) allocating different networks to learn the different tasks. Task decompositions are encouraged by enforcing a competition among the networks constituting the architecture. As a result of the competition, different networks learn different training patterns and, thus, learn to compute different functions. The architecture was first presented in Jacobs, Jordan, Nowlan, and Hinton (1991) and combines earlier work on learning task decompositions in a modular architecture by Jacobs, Jordan, and Barto (1991) with the mixture models view of competitive learning advocated by Nowlan (1990) and Hinton and Nowlan (1990).

The architecture, which is illustrated in Figure 2, consists of two types of networks: expert networks and a gating network. The expert networks compete to learn the training patterns, and the gating network mediates this competition. Whereas the expert networks have an arbitrary connectivity, the gating network is restricted to have as many output units as there are expert networks, and the activations of these output units must be nonnegative and sum to 1. The output of the entire architecture, denoted y, is the linear combination of the experts' outputs:

$$y = \sum_{i=1}^{n} g_i \, y_i, \tag{1}$$

where y_i denotes the output of the ith expert network and g_i is the gating network output corresponding to the ith expert.

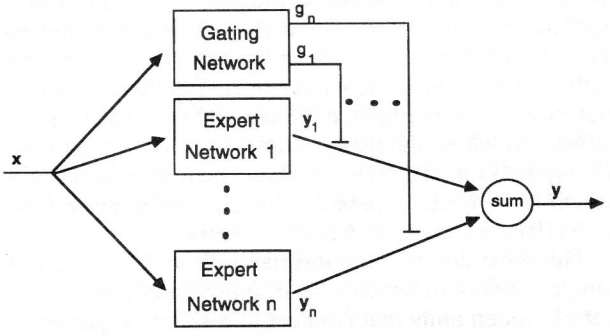

Figure 2. The mixtures-of-experts architecture consists of expert networks and a gating network. The expert networks compete to learn the training patterns; the gating network mediates the competition.

The learning process of the ME architecture combines aspects of competitive and associative learning. Mathematically, the architecture may be characterized as a probability model known as a conditional mixture density model. The form of the mixture components depends on the nature of the task: For regression tasks, the model is a conditional mixture of normal distributions; for classification tasks, the model is a conditional mixture of binomial or multinomial distributions. Because the architecture is a probability model, it is possible to quantify its performance using an appropriate likelihood function. The architecture's learning process is an optimization process that attempts to maximize the likelihood function. A mathematical description may be found in Jacobs, Jordan, Nowlan, and Hinton (1991), Jacobs and Jordan (1993), Jordan and Jacobs (1994), and Peng, Jacobs, and Tanner (1996). Here I present an intuitive description. During training, the connection strengths of the expert and gating networks are adjusted simultaneously. Each expert network's output is compared with the target output at each time step. The expert whose output most closely matches the target is called the winner of the competition; the other experts are called losers. An expert receives an amount of training information that is proportional to its relative performance on the training pattern. Whereas the winning expert receives a lot of information, and thus learns a lot about the current training pattern, the losing experts receive little or no information, and thus learn little about the current pattern. The gating network receives a different type of training information than do the expert networks. The experts receive information about the current input–output pattern provided by the environment. In contrast, the gating network receives information about the relative performances of the experts on the current pattern. It adjusts its connection strengths so that when the current input (or a similar input) recurs in the future, the activation of its output unit corresponding to the winning expert will be larger (i.e., closer to 1) and the activations of its remaining output units will be smaller (closer to 0).

The learning process has a positive feedback effect that forces different expert networks to learn different

tasks. This effect relies on the fact that, in general, input–output training patterns from the same task share a common underlying structure, whereas patterns from different tasks have different underlying structures. Suppose that at some time step, an expert has won the competition to learn some of the training patterns from one particular task. The expert will, therefore, have at least partial "knowledge" of the structure of the task. Consequently, in the future it will be likely to win the competition for the remaining patterns from that task. The expert will thereby become specialized for performing the task. As a result of this specialization, however, this expert will be likely to perform poorly on patterns from other tasks—unless some tasks happen to be very similar. Thus other experts will be likely to win the competition for the patterns from other tasks. In this way, different experts win the competition to learn patterns from different tasks, and the experts become specialized for performing different tasks.

From the viewpoint of the cognitive neurosciences, an interesting property of the ME architecture is the roles it assigns to nature and nurture in the acquisition of functional specializations. A feature of the architecture is that it tends to allocate to each task an expert network whose structure is well matched to that task. Structural properties of a network, such as its topology or receptive field characteristics, bias a network so as to make it a particularly good learner for some tasks but a poor learner for other tasks. Note that it is not simply the case that more complex networks (e.g., networks with many units, all connected to each other) learn faster than simpler networks. Instead, there exist structure–function correspondences, meaning that the structural properties of a network influence the set of tasks that the network can learn quickly, and the set that it can learn only with difficulty, if at all. Although little is formally known about the relationships between a network's structure and the nature of a task, some simple cases are clearly understood. For example, linear networks (i.e., networks without hidden units) learn to perform linear tasks faster than nonlinear networks; however, they cannot learn to perform nonlinear tasks. When expert networks with different structural properties compete to learn the training patterns, each network tends to win the competition for those patterns belonging to the task for which its structure makes it a good learner. Consequently, the architecture is capable of discovering structure–function relationships. The performance of the architecture is consistent with the theory that genetic instructions do not necessarily stipulate directly the function to be performed by each region of the brain (e.g., there is no genetic stipulation that language processing is predominately performed by the left cerebral hemisphere). Instead, genetic instructions bias the acquisition of functional specializations by assigning different structural properties to different regions. These structurally different regions may then, due to their performance characteristics, take on particular functions for which they are well suited (see Bever, 1980, and Kosslyn, 1987, for related processing accounts of cerebral lateralization).

As noted, the gating network composes the output of the ME architecture in the sense that it determines the extent to which each expert's output contributes to the output of the architecture as a whole. Note, however, that the ME architecture is equivalent to another architecture that contains expert networks but does not contain a gating network (by equivalent I mean that the two systems are exact notational variants of each other). Instead, the new architecture contains inhibitory connections among the expert networks so that each expert can suppress the outputs of the other experts. The strengths of these inhibitory connections are context dependent because they depend on the value of the current input pattern (units whose connection strengths are context dependent are known as sigma-pi units, Rumelhart, Hinton, & McClelland, 1986). At the end of training, the expert that was the winner of the competition in the context of the current input (or closely similar inputs) strongly suppresses the outputs of the other experts; experts that were losers of the competition do not suppress, or only weakly suppress, the other experts' outputs. This new architecture is notable in part because it highlights the fact that if one attempts to find a literal correspondence between neural systems and the ME architecture, there is no need to speculate about which specific structure in human nervous systems might correspond to the gating network. Instead, the job of the gating network can be performed by sets of inhibitory connections among neural modules. Further, this architecture is notable because the results of some experiments may be interpreted as suggesting that neural modules may use inhibitory interactions of this kind.

For example, Gazzaniga (1977; cited in Glass, Holyoak, & Santa, 1979) presented different visual inputs simultaneously to each hemisphere of a split-brain patient (i.e., a patient who has had the corpus callosum severed; this structure normally carries signals between the two cerebral hemispheres). While the patient centered his/her gaze on a fixation point, a word was briefly presented so that half of the letters fell to one side of the point and half the letters fell to the other side. If, for instance, the word was *target*, then *tar* fell in the patient's left visual field and was processed by the right hemisphere, whereas *get* appeared in the right visual field and was processed by the left hemisphere. When presented with four alternatives and asked to point to the one that matched the visual input, the patient consistently pointed to the stimulus that was presented to the right visual field regardless of which hand the patient used to perform the task. According to Glass et al. (1979), these results indicate that when there is a conflict between two plausible responses in this task, the left hemisphere inhibits the outputs of the right hemisphere and assumes motor control of both the left and right hands in making the response. As a second example, the hypothesis that the left hemisphere normally inhibits the right hemisphere during the

performance of linguistic tasks can also be found in the functional localization model of Moscovitch (1973). He speculated that the linguistic competence of the right hemisphere of normals is equal to that of split-brain subjects. The reason why experimental results often show that the right hemisphere of normals has limited linguistic abilities is that the left hemisphere normally inhibits the right hemisphere's attempts to process information linguistically through inhibitory influences across the corpus callosum. Experimental and theoretical results of this sort suggest that the inhibitory interactions among the expert networks of a ME architecture, as notationally embodied in the architecture's gating network, may closely resemble some types of interactions among neural modules.

SIMULATION RESULTS

The ME architecture can be used for many different purposes. Some of my research has concerned the characteristics of specialization or modularity in general; other parts of my work have concerned particular hypotheses about high-level visual processing. For example, the ME architecture can be used to compare the relative efficiencies with which different proposed sets of specializations can be acquired, and can also be used to compare the efficiency of learning in a device that acquires functional specializations versus that of single network devices that do not perform task decompositions. In addition, because the ME architecture tends to allocate to each task an expert network whose structure is well matched to that task, it is capable of discovering the structure–function relationships that are suitable for particular cognitive functions.

Jacobs, Jordan, Nowlan, and Hinton (1991) illustrated the use of an ME architecture on a spoken vowel discrimination task. The data were collected by Peterson and Barney (1952) and consisted of the first two formants of vowel utterances spoken by 75 different speakers. Exemplars from four vowel classes are shown in Figure 3 (vowels [i], [ɪ], [a], and [ɑ]). The horizontal axis of this figure gives the first formant value of each utterance, and the vertical axis gives the second formant value (the formant values have been linearly scaled). The lines within the graph in the figure show a typical final outcome of training an ME architecture. The line labeled Net 1 indicates that the system used one expert network to discriminate between instances of the vowels [i] and [ɪ], whereas the lines labeled Net 0 and Net 2 indicate that two experts were used to discriminate between the vowels [a] and [ɑ]. The boundary between the instances of [a] and [ɑ] has a slight bend in it, and so the architecture approximated this boundary by using different experts on each side of the bend. The line labeled Gate 0:2 indicates that the gating network turned on Expert Network 0 (i.e., $g_0 \approx 1$) for instances to the left of this line, whereas it turned on Expert 2 for instances to the right of the line. For our purposes, this simulation serves to illustrate two points. First, at the end of training, expert networks are functionally specialized. For example, Expert Network 1 is able to discriminate between the vowels [i] and [ɪ], but it has not learned anything about the vowels [a] and [ɑ]. Second, different expert networks become specialized for different functions; for example, Expert 1 has a different functional specialization than do Experts 0 and 2. A drawback of this simulation for current purposes is that it does not illustrate the fact that the ME architecture is sensitive to structure–function correspondences (the networks in this simulation were identical except for the initial random settings of their connection strengths). However, the three simulations described below highlight this property.

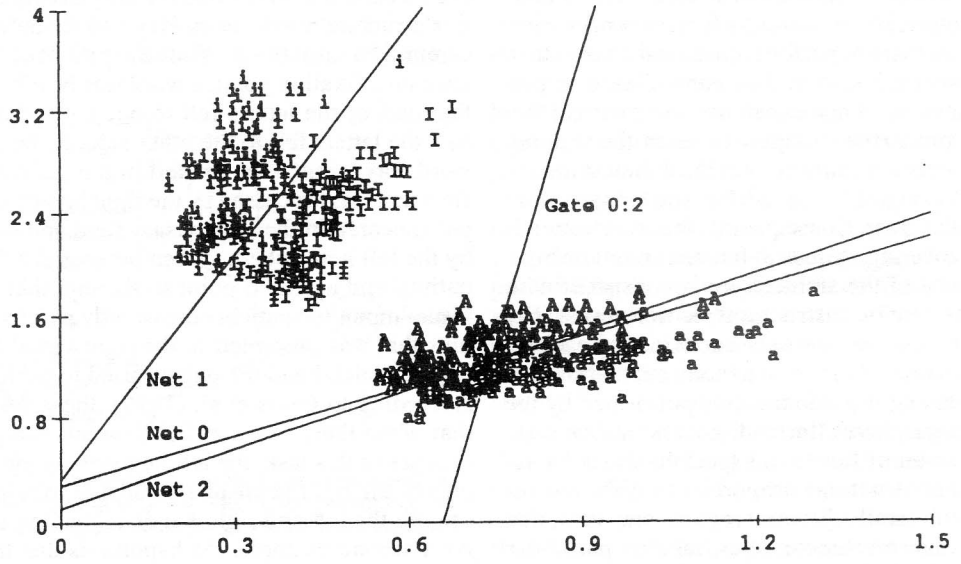

Figure 3. Data for the vowel discrimination task, and expert and gating networks decision lines at the end of training.

Jacobs, Jordan, and Barto (1991) provided an illustration of how the ME architecture may allocate a network with an appropriate structure to each of several high-level visual tasks. An architecture was trained to perform a visual object recognition task and a spatial localization task (these tasks were designed by Rueckl et al., 1989, and are meant to be analogous to the "what" and "where" visual functions attributed to temporal and parietal cortical pathways, respectively; see Mishkin, Ungerleider, & Macko, 1983). The data were formed by placing one of nine different objects at one of nine different locations on a retinal array. The localization task (identify the object's retinal location), but not the recognition task (identify the object on the retina), is linearly separable, meaning that it can be performed by a network with no hidden units. The ME architecture consisted of three expert networks, each with a different structure. The first expert had 36 hidden units, the second expert had 18 hidden units, and the third expert did not have any hidden units. Although either of the expert networks that contained hidden units could have learned to perform both tasks, that is not how the ME architecture allocated its expert networks. Rather, the architecture tended to use the expert with no hidden units to learn the localization task, and a network with hidden units to learn the recognition task. By doing so, the architecture showed an appropriate match between the structure of its expert networks and the nature of the tasks; a network without hidden units won the competition to learn the linearly separable localization task, whereas a network with hidden units won the competition to learn the nonlinear recognition task.

As a different example, Jacobs and Kosslyn (1994) considered the hypothesis that different subsystems of the brain are responsible for making categorical visual judgments and for making coordinate visual judgments. Categorical judgments include classifying the spatial relations between two stimuli (e.g., Object A is above/below Object B) and classifying the identity of a stimulus (e.g., Object A is a dog). Coordinate judgments include evaluating quantified spatial relations (e.g., Object A is 3.5 in. away from Object B) and identifying a visual stimulus as a particular exemplar (e.g., Object A is Fido). Note that much of the information needed to make categorical judgments is irrelevant for making coordinate judgments and, conversely, much of the information needed to make coordinate judgments is irrelevant for making categorical judgments. Categorization, for example, requires that various exemplars be grouped and treated as equivalent, whereas the identification of individual exemplars requires treating the instances as distinct. Therefore, from an information processing viewpoint, it is logical that the brain might use different subsystems to make categorical and coordinate visual judgments. Jacobs and Kosslyn reviewed experimental evidence from normal subjects for a double dissociation between categorical and coordinate judgment tasks. Laeng (1994) found the same double dissociation for categorical and coordinate spatial relations judgments in a study using unilateral stroke patients.

Kosslyn, Chabris, Marsolek, and Koenig (1992) speculated that there might be a structure–function relationship between receptive field sizes and visual judgments. Systems that make categorical visual judgments should be more efficient if they monitor visual neurons with small, nonoverlapping receptive fields (populations of such neurons provide relatively low-resolution representations of visual images), whereas systems that make coordinate visual judgments should be more efficient if they monitor neurons with large, overlapping receptive fields (populations of such neurons provide high-resolution representations of visual images).[1] Cowin and Hellige (1994) provided experimental evidence in support of this relationship by examining the effects of dioptric blurring on the performance of different spatial processing tasks using the same visual stimuli. Jacobs and Kosslyn (1994) used computer simulations to evaluate the proposed structure–function relationship, training neural networks to identify each visual stimulus as a member of a particular category ("shape category task") or to identify a stimulus as a particular exemplar ("shape coordinate task"). Networks did not view the visual stimuli directly; the stimuli were filtered through Gaussian units with restricted receptive fields. Figure 4 shows the units' receptive fields superimposed on top of the input array. The units with small receptive fields (left) provide a low-resolution representation of the array, whereas the units with large receptive fields provide a high-resolution representation. It was found that, indeed, the category task was learned faster when the receptive fields were relatively small, whereas the coordinate task was

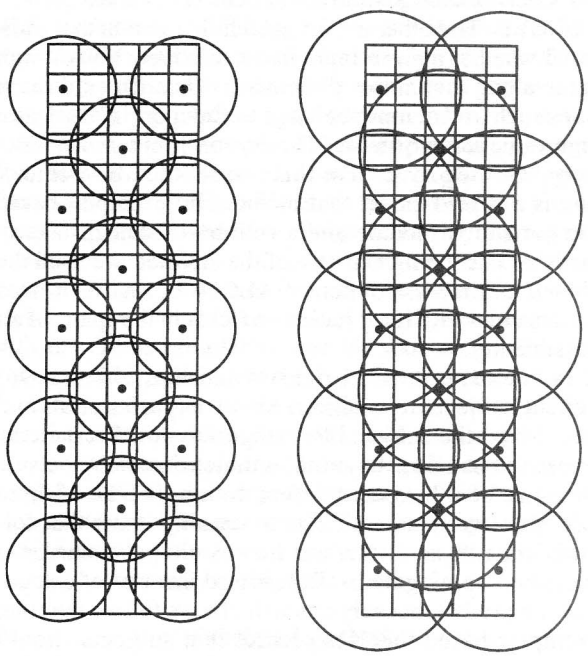

Figure 4. The Gaussian units' receptive fields superimposed on the input array. The units with small receptive fields (left) provide a low-resolution representation of the array, whereas the units with large receptive fields provide a high-resolution representation.

learned faster when the receptive fields were relatively large. Networks in which the receptive field sizes were allowed to adapt developed small receptive fields when trained on the categorical task and large receptive fields when trained on the coordinate task. When using an ME architecture, expert networks with small receptive fields tended to win the competition for the category task, whereas the coordinate task tended to be won by experts with large receptive fields. Overall, this set of simulations supports the hypothesized set of functional specializations and structure–function relationships by showing that these specializations and relationships are computationally efficient.

A third illustration of the ME architecture's sensitivity to structure–function relationships was provided by Erickson and Kruschke (1996; see also Kruschke & Erickson, 1994). These authors conducted a number of experimental studies to assess how people learn to categorize a set of visual items and to test how people use their knowledge of the categories when evaluating novel stimuli. Most of the training and test items used in the studies could be categorized according to a rule, though there also existed some items that were exceptions to the rule. The investigators compared the experimental results to the results of training two computational architectures. The first architecture was ALCOVE (Kruschke, 1992), an exemplar-based neural network that has been fit successfully to a wide range of data from categorization experiments. In short, ALCOVE uses hidden units whose receptive fields are restricted, local regions in the space of possible visual inputs. The second architecture was an ME architecture with two expert networks. These experts networks had different structures. One expert was ALCOVE. The other expert used hidden units that indicated whether a given input had a relatively low or high value along a stimulus dimension. Thus it could learn rules such as "an input belongs to Category A if it has a high value along Stimulus Dimension 1; otherwise it belongs to Category B." For this reason, the ME architecture is a hybrid model that includes an exemplar-based categorization module and a rule-based categorization module. The gating network of the architecture used the hidden unit representation of ALCOVE; that is, it used hidden units with local receptive fields in the space of all possible inputs.

In one experiment, Erickson and Kruschke (1996) trained subjects to categorize a variety of stimuli and then asked the subjects to categorize novel test items whose values along the stimulus dimensions were outside the range of values used during training. Some of these test items were more similar to training items that followed the rule, and other test items were more similar to exception training items. Rule-based theories of categorization predict that subjects will always follow the rule; exemplar-based theories predict that subjects should treat as exceptions those test items that are most similar to exception training items. The empirical results support the rule-based theories. Comparison of the performance of the computational models revealed that the ME architecture's performance agreed with the empirical data, but ALCOVE's did not.

In a second experiment, Erickson and Kruschke (1996) varied the frequency with which both rule and exception training items were presented. Results suggest that subjects tended to overgeneralize during early periods of training; that is, they tended to treat exception stimuli as rule stimuli. The ME architecture, but not ALCOVE, showed qualitatively similar behavior. An analysis of the test phase of the experiment showed that training item frequencies influenced subjects' classifications of test stimuli. Subjects showed fewer rule responses to test stimuli when the frequency of similar exception training items was increased and showed more rule responses when the frequency of similar rule training items was increased. Once again, the ME architecture, but not ALCOVE, also showed this behavior. Erickson and Kruschke concluded that because the ME architecture contains two different expert networks, one whose structure facilitated exemplar-based processing and one whose structure facilitated rule-based processing, it could account for the major findings in the experimental data, whereas the exemplar-based ALCOVE model could not.

VARIANTS AND EXTENSIONS OF THE ME ARCHITECTURE

The ME architecture is one member of a family of architectures. The importance of other members arises when one considers additional learning theory issues in the context of modular systems. For example, complex tasks may often be decomposed into subtasks that are only roughly distinct; that is, the subtasks may share some common features. The ME with a share network is an architecture that uses one network (a share network) to learn the shared features and uses different networks (expert networks) to learn the distinct features of each subtask. As before, the experts compete to learn the training patterns. The share network does not take part in the competition; it attempts to learn features of all training patterns. The output of the architecture, $\mathbf{y}$, is the sum of the output of the share network, $\mathbf{y}_s$, and the gated outputs of the expert networks:

$$\mathbf{y} = \mathbf{y}_s + \sum_{i=1}^{n} g_i \mathbf{y}_i. \quad (2)$$

The training procedure for this architecture is identical to the training procedure for the standard ME architecture with the exception that the networks' weights are adjusted so as to maximize a likelihood function that takes into account the presence of the share network.

In addition to the study of high-level vision, the ME family of architectures has also been applied to the adap-

tive control of dynamical systems. The dynamics of nonlinear systems often vary qualitatively over their parameter space. Methods for designing piecewise control laws for dynamical systems, such as gain scheduling, are useful because they circumvent the problem of determining a single global model of the system dynamics. Instead, the system dynamics are approximated using local models that vary with the system's operating conditions. When a controller is learned instead of designed, analogous issues arise. The standard ME architecture and the ME with a share network represent novel approaches to learning piecewise control laws. Jacobs and Jordan (1993) found that the ME with a share network showed fast learning of the inverse dynamics of a simulated two-joint robot arm that was required to move a variety of payloads, each of a different mass, along a desired trajectory. The share network learned to supply the torques necessary to control the robot arm with no payload, and different expert networks learned to add extra torques to compensate for the mass of different payloads. That is, one expert learned to add extra torques for light payloads, another expert added extra torques for payloads of moderate mass, and a third expert added extra torques for heavy payloads.

If it is useful to divide a task into subtasks, then it ought to be useful to divide subtasks into subsubtasks, and so on. The ME architecture can be extended to a hierarchical mixtures-of-experts (HME) architecture that uses competition to recursively split the input space into nested regions and to learn separate associative mappings within each region. Figure 5 illustrates a two-level hierarchy. The first level (left) consists of two ME architectures. The outputs of these architectures are weighted by a gating network at the second level of the hierarchy (right). In general, the networks of the hierarchy may be trained to maximize an appropriate likelihood function via a hill-climbing procedure (Jordan & Jacobs, 1992). However, my colleagues and I have found it useful to pursue other optimization procedures when using an HME architecture.

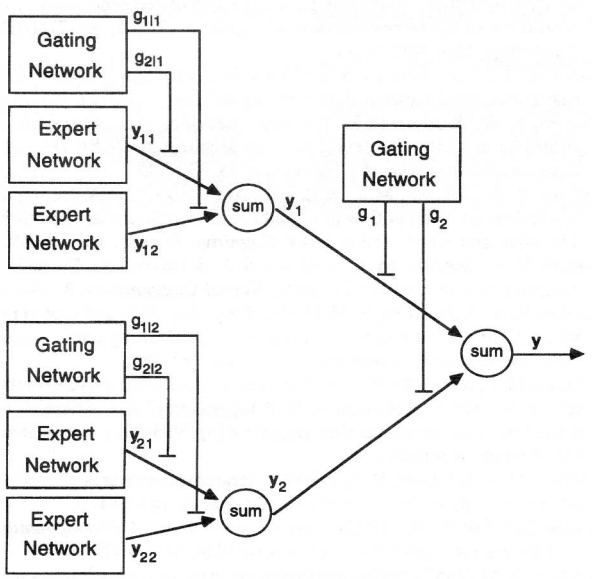

Figure 5. A two-level hierarchical mixtures-of-experts (ME) architecture. The outputs of two ME architectures at the lowest level (left) are weighted by a gating network at the highest level (right).

An important theoretical question concerning learning in modular systems is whether or not adaptation strategies are available to modular systems that are not available (or not easily available) to nonmodular devices that allow the modular systems to show rapid learning. If we restrict ourselves to a special case of the HME architecture, namely the case in which each expert and gating network is a generalized linear model (i.e., a model consisting of a linear transformation followed by a monotone nonlinear transformation so that the distribution of the model's output is a member of the exponential family), then it appears that such a strategy exists. The HME architecture may be trained to maximize a likelihood function using the expectation-maximization (EM) algorithm (Jordan & Jacobs, 1994). Nonlinear learning systems that lack a modular structure often cannot take advantage of this algorithm. The EM algorithm is an iterative, non-gradient–based algorithm for maximizing likelihood functions that has proven to be extremely efficient in a wide variety of applications (Dempster, Laird, & Rubin, 1977). Indeed, the HME architecture trained with the EM algorithm learned roughly two orders of magnitude faster than a single network trained with a hill-climbing algorithm on a relatively difficult robot dynamics task (Jordan & Jacobs, 1994).

Peng et al. (1996) used a different algorithm to train an HME architecture. This algorithm is a Bayesian inference algorithm known as a Gibbs sampler. In general, an advantage of Bayesian sampling techniques is that they do not simply provide a point estimate of the expected value of a random variable, but rather provide the entire distribution of the variable. For present purposes, this means that an HME architecture trained via the Gibbs sampler produces both an estimate of the correct response and a measure of how confident it is in its estimate. This confidence measure is useful for obtaining robust performance. Peng et al. trained an HME architecture with the Gibbs sampler to classify spoken vowels. The inputs to the system were the first two formants of each utterance; the target output was the correct vowel classification. After training, the HME architecture produced the vowel class to which it thought each utterance belonged, and it also produced a measure of confidence in each of its classification predictions. Appropriately, the architecture had the lowest confidence in its classifications of acoustically ambiguous utterances.

CONCLUSIONS

In summary, this article has noted the increasing evidence that experience-dependent processes are more influential in determining a brain region's structural and

functional properties than was previously supposed. Consequently, one may study the developmental principles that play a role in the acquisition of functional specializations. My research program studies the hypothesis that a combination of structure–function correspondences plus the use of competition between modules leads to functional specializations. This principle has been instantiated in a family of computational architectures referred to as ME architectures, and in a corresponding family of learning rules. The article has reviewed a number of ME architectures, including a single-level modular system, a modular system in which some, but not all, of the networks participate in the competition, and a multilevel hierarchical system. A variety of learning rules have also been discussed, including a hill-climbing procedure, a statistical procedure known as the EM algorithm that leads to rapid learning, and a Bayesian procedure that allows for the computation of confidence measures. Simulation results suggest that an important feature of ME architectures is that they are sensitive to structure–function correspondences. Each network tends to win the competition to learn the tasks for which its internal structure makes it a particularly good learner.

The approach proposed here may be viewed as advocating a middle ground in the nature versus nurture debate. In this sense it is compatible with the views of a growing number of developmental psychologists. Karmiloff-Smith (1992), for example, contrasted the views of nativists (e.g., Fodor, 1983), with those of constructivists (e.g., Piaget, 1955). Whereas nativists emphasize the existence of built-in, domain-specific knowledge, and domain-specific processing modules, constructivists stress a minimal innate underpinning to subsequent domain-general learning. Karmiloff-Smith argued that these two seemingly contradictory views can be reconciled if one posits that modularization, or functional specialization, develops gradually over time in a way that is shaped by an organism's experiences. The function acquired by each module is biased but not strictly determined by genetic factors.

A long-term goal of researchers is to delineate how such genetic factors act as a bias. If competition plays a role in the acquisition of functional specializations in the way that has been proposed here, it is clear that there must exist initial differences between the competing modules. The distinctive features of a module make it a comparatively good learner for some tasks, but a poor learner for other tasks. Other modules, with different distinctive features, have different learning biases. One way in which modules may initially differ is in the information carried by the inputs they receive. A module primarily receiving visual inputs will perform better on a visual task than a module primarily receiving auditory inputs. Results with the ME architecture suggest the importance of other differences between modules. Differences in receptive field size, number of processing units, and connectivity among the processing units all serve to bias a module's learning performance, thereby biasing its functional specialization.

REFERENCES

Ballard, D. H. (1986). Cortical connections and parallel processing: Structure and function. *Behavioral & Brain Sciences*, **9**, 67-120.

Bever, T. G. (1980). Broca and Lashley were right: Cerebral dominance is an accident of growth. In D. Kaplan & N. Chomsky (Eds.), *Biology and language* (pp. 186-230). Cambridge, MA: MIT Press.

Cowin, E. L., & Hellige, J. B. (1994). Categorical versus coordinate spatial processing: Effects of blurring and hemispheric asymmetry. *Journal of Cognitive Neuroscience*, **6**, 156-164.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**, 1-38.

Dennis, M., & Kohn, B. (1975). Comprehension of syntax in infantile hemiplegics after cerebral hemidecortication: Left-hemisphere superiority. *Brain & Language*, **2**, 472-482.

Dennis, M., & Whitaker, H. A. (1976). Language acquisition following hemidecortication: Linguistic superiority of the left over the right hemisphere. *Brain & Language*, **3**, 404-433.

Erickson, M. A., & Kruschke, J. K. (1996). *Modeling human performance in a categorization task with rules and exceptions: The importance of interaction*. Poster presented at the 18th Annual Conference of the Cognitive Science Society. Available http://www.indiana.edu/~kruschke/home.html

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Gazzaniga, M. S. (1977). Consistency and diversity in brain organization. In S. J. Dimond & D. A. Blizard (Eds.), *Evolution and lateralization of the brain* (Annals of the New York Academy of Sciences, Vol. 299, pp. 415-423). New York: New York Academy of Sciences.

Geschwind, N., & Galaburda, A. M. (1987). *Cerebral lateralization: Biological mechanisms, associations, and pathology*. Cambridge, MA: MIT Press.

Glass, A. L., Holyoak, K. J., & Santa, J. L. (1979). *Cognition*. Reading, MA: Addison-Wesley.

Hinton, G. E. (1981). Shape representation in parallel systems. In A. Drina (Ed.), *Proceedings of the Seventh International Joint Conference on Artificial Intelligence* (pp. 1088-1096).

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 77-109). Cambridge, MA: MIT Press.

Hinton, G. E., & Nowlan, S. J. (1990). The bootstrap Widrow–Hoff rule as a cluster-formation algorithm. *Neural Computation*, **2**, 355-362.

Jacobs, R. A., & Jordan, M. I. (1993). Learning piecewise control strategies in a modular neural network architecture. *IEEE Transactions on Systems, Man, & Cybernetics*, **23**, 337-345.

Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science*, **15**, 219-250.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, **3**, 79-87.

Jacobs, R. A., & Kosslyn, S. M. (1994). Encoding shape and spatial relations: The role of receptive field size in coordinating complementary representations. *Cognitive Science*, **18**, 361-386.

Jordan, M. I., & Jacobs, R. A. (1992). Hierarchies of adaptive experts. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), *Advances in neural information processing systems 4* (pp. 985-992). San Mateo, CA: Morgan Kaufmann.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**, 181-214.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press.

Kosslyn, S. M. (1987). Seeing and imagining in the cerebral hemispheres: A computational approach. *Psychological Review*, **94**, 148-175.

Kosslyn, S. M., Chabris, C. F., Marsolek, C. J., & Koenig, O. (1992).

Categorical versus coordinate spatial representations: Computational analyses and computer simulations. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 562-577.

KRUSCHKE, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.

KRUSCHKE, J. K., & ERICKSON, M. A. (1994). Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model. In A. Ram & K. Eiselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 514-519). Hillsdale, NJ: Erlbaum.

LAENG, B. (1994). Lateralization of categorical and coordinate spatial functions: A study of unilateral stroke patients. *Journal of Cognitive Neuroscience*, **6**, 189-203.

MILNER, P. M. (1974). A model for visual shape recognition. *Psychological Review*, **81**, 521-535.

MISHKIN, M., UNGERLEIDER, L. G., & MACKO, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends In Neurosciences*, **6**, 414-417.

MOSCOVITCH, M. (1973). Language and the cerebral hemispheres: Reaction-time studies and their implications for models of cerebral dominance. In P. Pliner, L. Krames, & T. Alloway (Eds.), *Communication and affect: Language and thought* (pp. 89-126). New York: Academic Press.

NEVILLE, H. J. (1995). Developmental specificity in neurocognitive development in humans. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 219-231). Cambridge, MA: MIT Press.

NOWLAN, S. J. (1990). Maximum likelihood competitive learning. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2* (pp. 574-582). San Mateo, CA: Morgan Kaufmann.

OJEMANN, G., OJEMANN, J., LETTICH, E., & BERGER, M. (1989). Cortical language localization in left, dominant hemisphere. *Journal of Neurosurgery*, **71**, 316-326.

O'LEARY, D. D. M. (1989). Do cortical areas emerge from a protocortex? *Trends in Neurosciences*, **12**, 400-406.

PENG, F., JACOBS, R. A., & TANNER, M. A. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, **91**, 953-960.

PETERSON, G. E., & BARNEY, H. L. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

PIAGET, J. (1955). *The child's construction of reality*. London: Routledge & Kegan Paul.

PINKER, S., & PRINCE, A. (1988). On language and connectionism. *Cognition*, **28**, 73-194.

QUINLAN, P. (1991). *Connectionism and psychology*. Chicago: University of Chicago Press.

RUECKL, J. G., CAVE, K. R., & KOSSLYN, S. M. (1989). Why are "what" and "where" processed by separate cortical visual systems? A computational investigation. *Journal of Cognitive Neuroscience*, **1**, 171-186.

RUMELHART, D. E., HINTON, G. E., & MCCLELLAND, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 45-76). Cambridge, MA: MIT Press.

RUMELHART, D. E., HINTON, G. E. & WILLIAMS, R. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.

RUMELHART, D. E., MCCLELLAND, J. L., & THE PDP RESEARCH GROUP (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

SUR, M., PALLAS, S. L., & ROE, A. W. (1990). Cross-modal plasticity in cortical development: Differentiation and specification of sensory neocortex. *Trends in Neurosciences*, **13**, 227-233.

NOTE

1. Intuitively, the relationship between neurons' receptive field sizes and the resolution of the representation provided by these neurons can be understood by considering the overlap among the receptive fields. Suppose that a point of light appears on a retina, and that there exists a population of two visual neurons in which each neuron has a circular receptive field on the retina. Also assume that the two neurons' receptive fields touch at a point and so do not overlap. Because a neuron becomes active only when light falls within its receptive field, it is possible to distinguish three spatial locations on the retina with this population: Either both neurons are inactive, meaning that the point of light is outside either neuron's receptive field, or one of the two neurons is active, meaning that the point of light falls within the active neuron's receptive field. Now contrast this situation with a new one in which the neurons' receptive fields are 20% larger in diameter than they were in the first situation, so that there is some overlap among these fields, although not complete overlap. In this case, it is possible to distinguish four spatial locations on the retina: The light is outside either neuron's receptive field (both neurons are inactive), the light is in one neuron's receptive field but not the other's (one neuron is active and the other is inactive), or the light falls within the intersection of the neurons' receptive fields (both neurons are active). Because the neurons in the first case can be used only to distinguish three spatial locations, whereas the neurons in the second case can be used to distinguish four locations, it is said that the former neurons provide a relatively low-resolution representation and the latter neurons provide a high-resolution representation. For present purposes, it is possible to extrapolate from these two situations in order to conclude that populations of neurons with relatively small, nonoverlapping receptive fields can provide a low-resolution representation, whereas populations of neurons with large, overlapping receptive fields can provide a high-resolution representation. See Ballard (1986), Hinton (1981), Hinton, McClelland, and Rumelhart (1986), and Milner (1974) for further discussions of this point.

(Manuscript received August 12, 1996;
revision accepted for publication January 21, 1997.)