Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 12370–12373, December 1995
Neurobiology

Brain localization for arbitrary stimulus categories: A simple account based on Hebbian learning

THAD A. POLK* AND MARTHA J. FARAH

Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104

Communicated by David E. Rumelhart, Stanford University, Stanford, CA, September 8, 1995

ABSTRACT **A central theme of cognitive neuroscience is that different parts of the brain perform different functions. Recent evidence from neuropsychology suggests that even the processing of arbitrary stimulus categories that are defined solely by cultural conventions (e.g., letters versus digits) can become spatially segregated in the cerebral cortex. How could the processing of stimulus categories that are not innate and that have no inherent structural differences become segregated? We propose that the temporal clustering of stimuli from a given category interacts with Hebbian learning to lead to functional localization. Neural network simulations bear out this hypothesis.**

Localization of function is a basic feature of brain organization, revealed by the selectivity of impairments following brain damage and by techniques for recording regional brain activity. For many functions, such as color vision, motor control, and even face recognition, one can hypothesize that some combination of genetic factors and intrinsic differences in the stimuli themselves causes spatial segregation. However, neuropsychology provides evidence of localization for the processing of arbitrary stimulus categories, such as letters versus digits. A genetic account is not possible, as letters and digits entered the environment far too recently to be represented in the genome. Neither can the physical features of letters and digits explain their cortical segregation, as many letter–digit pairs are physically more similar than many letter–letter and digit–digit pairs.

Letters and digits are not the only categories that pose this dilemma. In addition to impairments in letter recognition relative to digit recognition (1), brain-damaged patients can show selective impairments in the processing of music relative to other sounds (3), in writing relative to other sensorimotor functions of the hand (4), and even in cursive relative to print (M. Kinsbourne and B. Hiltbrunner, personal communication). Intraoperative cortical stimulation suggests that different cortical regions subserve different languages in bilingual individuals (5). In all these cases, the impaired abilities are too recent in evolutionary terms to admit genetic explanations and yet there are no obvious, inherent differences in the required processing that could account for the differentiation. How might such stimulus categories come to be processed by different brain regions?

The answer may lie in the statistics of the environment. For all these categories, stimuli tend to co-occur in close temporal proximity. Letters more often appear with other letters than with numbers and vice versa. Similarly, musical sounds are more often followed by other musical sounds than by nonmusical sounds. If one has just written a letter, one is more likely to write another than to do something else with one's hand, and the same temporal clustering applies to writing cursive versus print. Finally, the elements of one language are more likely to be spoken and heard in close temporal proximity with one another than with those of another language. Given the correlation-driven nature of Hebbian learning, we hypothesized that this statistical property of the world could cause neural networks to self-organize spatially segregated representations for stimuli from such otherwise arbitrary categories.

How could the cortex pick up on this correlation in the environment to produce segregated representations for arbitrary categories such as letters and digits? Fig. 1 presents a simple mechanistic model that demonstrates one possibility. The model is a two-layer neural network that uses a Hebbian learning rule to modify the weights of the connections between the input and output layers. Hebbian learning is a neurophysiologically plausible mechanism that generally corresponds to the following rule: If two units are both firing (correlated), then their connection is strengthened; if only one unit of a pair is firing (anticorrelated), then their connection is weakened (6). The input layer represents the visual forms of input characters (letters and digits) by using a localist representation (each unit represents a different visual form). Initially, the output layer does not represent anything (since the connections from the input layer are initially random), but with training it should self-organize to represent letters and digits in segregated areas. Neighboring units in the output layer are connected via excitatory connections and units further away are connected via inhibitory connections, in keeping with previous models of cortical self-organization (7, 8). (Other architectures would also be consistent with our explanation—e.g., normalization of output activations as opposed to long-range inhibitory connections. What is critical is that the architecture provide a cooperative mechanism to produce clusters of activity and a competitive mechanism to inhibit multiple clusters. For a review of a variety of such models see ref. 9.) The legend to Fig. 1 describes the model's details.

Fig. 1 shows the state of the network at different points during training with letters and digits. When the first stimulus ("A") is initially presented, the pattern of output activity is random, reflecting the random initial connection strengths from inputs to outputs. Over time the excitatory connections produce clusters around active units and these drive down activity elsewhere via inhibitory connections, leading to a single cluster (or very few). The Hebb rule then strengthens the connections to this active cluster from the active input (A) but weakens the connections from other (inactive) inputs and those from A to inactive outputs because they are anticorrelated. Because of these weight changes, A will subsequently be biased toward activating the same cluster whereas other inputs will be biased toward activating other units. When another stimulus ("1") is presented, it too leads to a random pattern of activity, except that the units activated by A are less active than they otherwise would be. Consequently, the "1" cluster that develops is spatially segregated from the "A" cluster. Similarly, the clusters for "H" and "8" are spatially segregated from the others.

Fig. 2 shows the network's behavior when multiple characters are presented simultaneously. Here clusters form for

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*To whom reprint requests should be addressed.

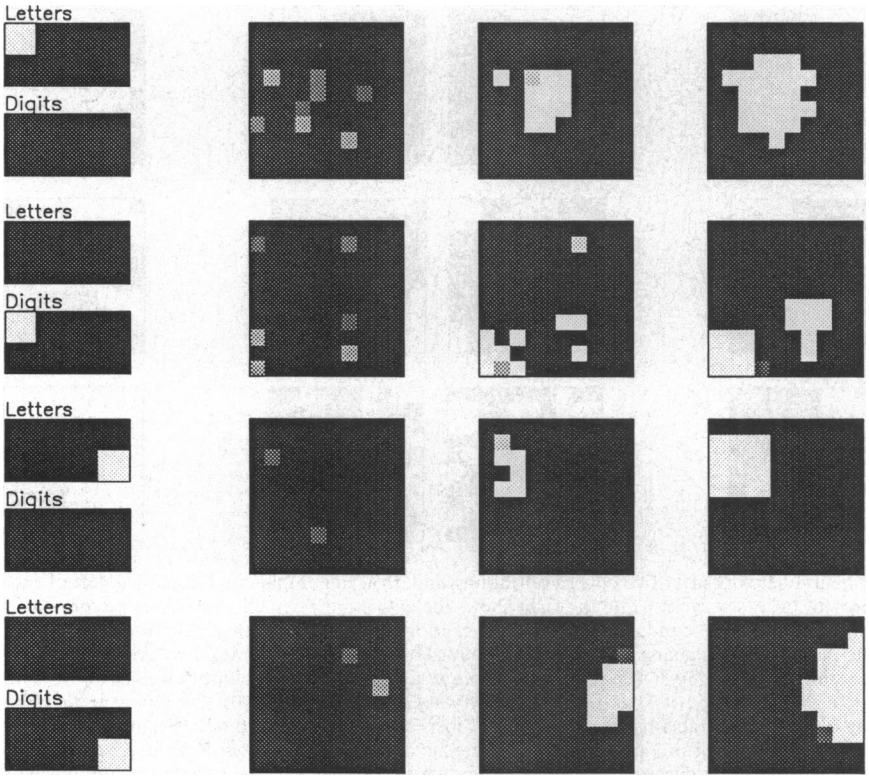

FIG. 1. The state of the network at different points during training with individual characters. Each row represents the presentation of a stimulus during training (only a few illustrative stimuli are shown). The stimulus is on the left. The rest of the row shows the state of the output units after 1, 3, and 5 cycles. The network had 116 total units. Sixteen were inputs (the first 8 for the letters A–H and the other 8 for the digits 1–8) and the other 100 were outputs in a 10×10 arrangement. All input units were connected to all output units with plastic connections. Output units were connected to neighbors (in two dimensions) by fixed excitatory connections (weight = 0.3) and to other output units by fixed inhibitory connections (weight = −0.03). The minimum and maximum unit firing rates were fixed at 0.0 and 100.0, while the minimum and maximum connection weights were fixed at 0.0 and 3.0. Initially, the activity of output units was uniform random between 0.0 and 10.0 and the connection weights from inputs to outputs were uniformly random between 0.0 and 0.5. The following Hebbian learning rule based on firing rate was used after every cycle to update connection strengths between input and output units: if both pre- and postsynaptic units are firing above threshold (50.0), increase connection weight by 0.08; if both units are below threshold, make no change; otherwise, decrease the connection weight by 0.025. The output units used a sigmoid transfer function:

$$\text{output} = \frac{100.0}{1 + e^{-(\text{input} - 40.0)}}.$$

The total input to each output unit was multiplied by a 0.9 gain factor before passing through the transfer function. The input units were clamped to their values and did not decay.

character sets instead of for individual characters. If two stimuli initially activate widely separated clusters but then appear together, the initial clusters will compete with each other (via the inhibitory connections) in representing the pair. One cluster will eventually win out and Hebbian learning will strengthen the connections from *both* inputs to the victorious cluster. The result is that co-occurring stimuli will be biased toward exciting nearby units, even if they initially excited quite different sets of units. In other words, spatially localized areas will develop for stimuli that tend to co-occur (as we assume within-category stimuli do). Stimuli that occur in rapid succession could also become associated if some residual activation from the first stimulus occurs (10).

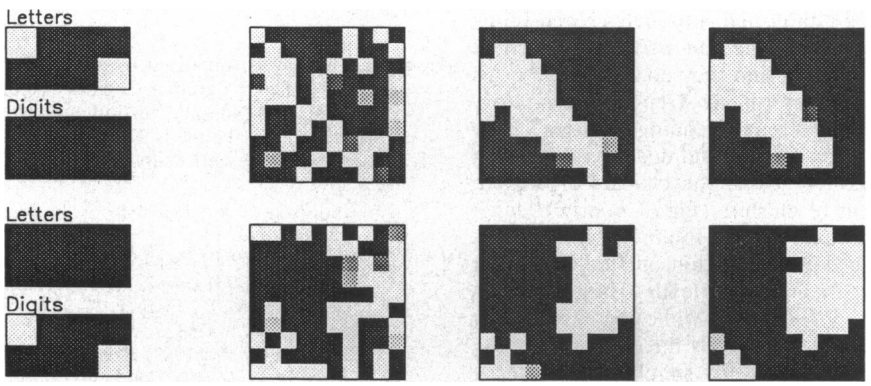

FIG. 2. The state of the network at different points during training with multiple characters. Conventions are the same as in Fig. 1.

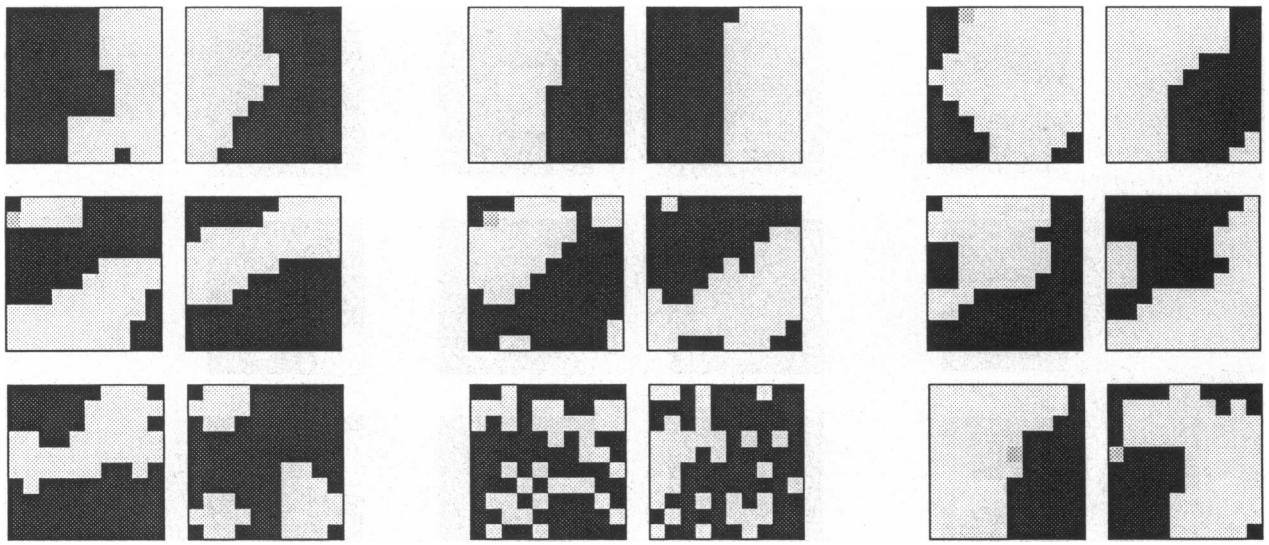

FIG. 3. The state of the neural network after five epochs of training under varying initial conditions. The left of each pair shows the activation when all eight letters are present (after six cycles), and the right shows the activation when all eight digits are present. The central pair shows the final state of a network that used the initial conditions given in the caption to Fig. 1. The other pairs used initial conditions that were identical except that the value of one parameter was changed. The pair directly above it (*Top Center*) used an initial excitatory connection weight of 0.5 between neighboring output units, instead of 0.3. The pair below (*Bottom Center*) used an excitatory weight of 0.1. The pair at *Top Right* used an inhibitory connection weight of −0.01 (instead of −0.03) and that at *Bottom Left* used one of −0.05. The pair at *Middle Right* used a different Hebbian rule in which the connection between simultaneously active units was increased by 0.1 (instead of 0.08) and the pair at *Middle Left* increased such connections by 0.06. Finally, if only one unit in a pair fired above threshold, the pair at *Bottom Right* decreased the connection strength by 0.01 (instead of 0.025), while the pair at *Top Left* decreased it by 0.04. Stimuli were presented in the following order during each epoch: A, 1, B, 2, . . ., H, 8 (16 single characters), AB, 12, BC, 23, . . ., HA, 81, AB, 12, . . ., HA, 81 (32 pairs of characters), ABC, 123, BCD, 234, . . ., HAB, 812, ABC, 123, . . ., HAB, 812 (32 sets of character triples). Each stimulus was presented for six cycles after which the output units were reset to 0.0 activation.

The same argument implies that stimuli that do *not* co-occur will be biased *away* from exciting nearby units. When one stimulus is present, the connections from any other (inactive) stimulus to the currently active output units will be weakened, and this will bias the inactive stimulus away from exciting those units. Consequently, stimuli from different categories (e.g., letters versus digits) will tend to be represented by spatially segregated sets of units (because we assume they co-occur much less frequently). And even *within* a category, as long as particular stimuli do not *always* co-occur, they will have distinct representations within their cortical areas.

Fig. 3 shows the results with different initial conditions. Distinct, spatially localized areas develop in almost every case. The parameters affect the size, coherence, and degree of overlap of clusters but do not change the qualitative pattern of results. Also note that the resulting letter and digit areas do not always form in the same locations. In some simulations, the letter area arises on the left of the output layer with the digit area on the right. In others, these locations are reversed, or horizontal or diagonal patterns arise. Whether the locations of letter and digit areas in human cortex vary across individuals in a similar way is not well established, although a recent study using electrodes chronically implanted on the surface of striate and extrastriate cortex suggests that they do (11).

The most important parameter is the strength of excitatory connections, with increasing strength leading to larger, more coherent clusters (Fig. 3 *Top Center*) and decreasing strength reducing the size and coherence of the clusters enough to undermine the formation of clusters (Fig. 3 *Bottom Center*; note the segregation in both cases). The inhibitory connections are important in driving down activation outside the major clusters. Thus, decreasing their strength leads to larger clusters with more overlap (Fig. 3 *Top Right*) while increasing their strength leads to smaller clusters with no overlap (*Bottom Left*). Although the learning rate for simultaneously firing units has no effect in the range we tried (*Middle* row), the learning rate for anticorrelated firing affects the degree of segregation. Increasing it biases different stimuli to activate different units and leads to greater segregation (*Top Left*), whereas decreasing it has the opposite effect (*Bottom Right*).

We have shown that some simple and familiar properties of self-organizing systems—namely, correlation-driven learning and short-range excitatory connections—in conjunction with a robust statistical feature of the environment, will lead to spatial localization for arbitrary stimulus categories. Principles of self-organization have previously been used to explain the development of map-like cortical representations, in which location within the network corresponds to stimulus location on the retina, skin, or cochlear membrane (12–19). Organizations reflecting the structure of more abstract spaces have also been obtained, including maps in which network location corresponds to location in a semantic space (e.g., the representations of "dog" and "cat" are closer together than those of "dog" and "table") (2). Unlike this semantic organization, the mechanism responsible for the localization of arbitrary categories involves simple, first-order stimulus statistics. It appears that the arbitrary categories for which there is evidence of cortical specialization occur in the world with just such statistics.

This work was supported by a grant-in-aid for training from the McDonnell–Pew Program in Cognitive Neuroscience to T.A.P. and by grants from the National Institutes of Health, the Office of Naval Research, the Alzheimer's Disease Association, and the Research Association of the University of Pennsylvania.

1. Anderson, S. W., Damasio, A. R. & Damasio, H. (1990) *Brain* **113,** 749–766.
2. Ritter, H. (1990) *Psychol. Res.* **52,** 128–136.
3. Peretz, I. (1993) *Cognit. Neuropsychol.* **10,** 21–56.
4. Alexander, M. P., Fischer, R. S. & Friedman, R. (1992) *Arch. Neurol.* **49,** 246–251.
5. Ojemann, G. A. (1983) *Behav. Brain Sci.* **6,** 189–230.
6. Hebb, D. O. (1949) *The Organization of Behavior: A Neuropsychological Theory* (Wiley, New York).
7. von der Malsburg, C. (1973) *Kybernetik* **14,** 85–100.

8. von der Malsburg, C. (1979) *Biol. Cybern.* **32,** 49–62.
9. Goodhill, G. (1992) Thesis (Univ. of Sussex, Brighton, U.K.).
10. Foldiak, P. (1991) *Neural Comput.* **3,** 194–200.
11. Allison, T., McCarthy, G., Nobre, A., Puce, A. & Belger, A. (1994) *Cereb. Cortex* **4,** 544–554.
12. Cottrell, M. & Fort, J. C. (1986) *Biol. Cybern.* **53,** 405–411.
13. Durbin, R. & Mitchison, G. (1990) *Nature (London)* **343,** 644–647.
14. Kohonen, T. (1982) *Biol. Cybern.* **43,** 59–69.
15. Kohonen, T. (1988) *Self-Organization and Associative Memory* (Springer, New York), 2nd Ed.
16. Linsker, R. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 7508–7512.
17. Linsker, R. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 8390–8394.
18. Linsker, R. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 8779–8783.
19. Miller, K. D., Keller, J. B. & Stryker, M. P. (1989) *Science* **245,** 605–615.