On Self-Organization in Connectionist Networks

Ronald J. Williams
Institute for Cognitive Science
University of California, San Diego C-015
La Jolla, CA 92093

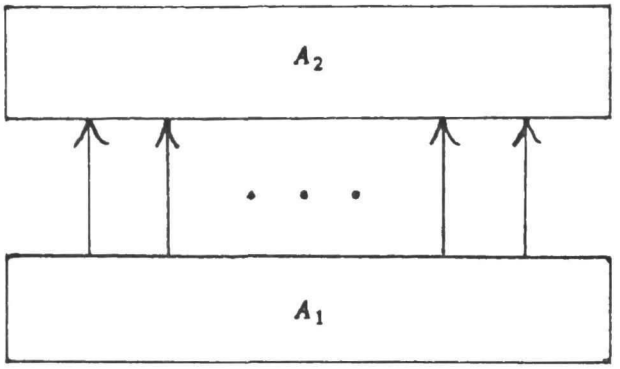
The aim of this paper is to present some observations about certain types of representations, or encodings, in connectionist, or neural-like, networks. In particular, this paper will call attention to two distinct categories of encoding in such networks and examine some results bearing on the issue of self-organizing networks which use one or the other type of encoding. This discussion will be limited to the encoding of data which is fundamentally numerical (or, more precisely, geometric). It is an interesting question whether semantic data can also be imbedded in a geometric framework, but such matters will be ignored here.

A number of interesting attempts have been made to provide an answer to the general problem of how a network might be shaped to a particular environment through self-organization. Among these are the early perceptron studies of Rosenblatt (1962), the investigations into possible neural net dynamics by Grossberg (1980), the recent theoretical approach of Hinton & Sejnowski (1983), and several works with the goal of finding ways in which cells in visual cortex might become tuned to specific features through self-organization (von der Malsburg, 1973; Nass & Cooper, 1975; Bienenstock et al., 1982). Two recent works which share a common perspective with the approach to be taken here are that of Kohonen (1982) and that of Amari (1983).

On the other hand, a number of non-self-organizing connectionist networks have been hand-crafted to perform particular sensory or cognitive processing tasks in ways which are generally intended to account for human performance data and/or be compatible with neurobiological data (Feldman & Ballard, 1982; Ballard, 1981; McClelland & Rumelhart, 1981; Hinton, 1981). Certain classes of network have even been proposed as having a certain universality in sensory processing, at least in the visual system (Ballard, 1981). Such universality might reasonably be taken as making such networks plausible candidates for the actual implementation of these algorithms in the brain. It then becomes reasonable as well to investigate possible mechanisms by which such networks might be able to self-organize to some degree; if such mechanisms can be shown to exist, it could then be argued that these types of network represent a general processing strategy which could find wide applicability in the brain.

This research was supported by a grant to David Zipser from the System Development Foundation.

In what follows, attention will be restricted to the following 2-layer architecture:



A_1 and A_2 are layers of units and each may optionally have fixed lateral connections. As depicted, there are connections from A_1 units to A_2 units, and these will be assumed to be variable and thus subject to self-organization. In addition, there may be connections (not depicted above) from A_2 units to A_1 units, and these may also be variable. The reason that only the $A_1 \rightarrow A_2$ connectivity is depicted above is that it is the $A_1 \rightarrow A_2$ transfer function that is of paramount interest here. Specifically, input to the system will be assumed to consist of a pattern of activation in the A_1 layer, and output will be taken as the resulting pattern of activation in the A_2 layer. While this network is being considered here in isolation, one may view this more generally as simply a sub-network consisting of two adjacent layers in a larger hierarchical network.

I will consider the $A_1 \rightarrow A_2$ transfer function as performing a mapping between two individual encodings, from that in the A_1 layer to that in the A_2 layer. A pattern in either layer can be considered as a Euclidean vector whose coordinates are simply the respective activation values of all of the units in that layer. The distinction I suggest drawing between encodings essentially revolves around how useful such a vector space description is for capturing the essential dimensions along which the lawful patterns may vary.

A full characterization of the two types of encoding will not be given here; it will be sufficient for present purposes to simply give examples of each and to cite a closely related distinction already existing in the literature.

Ballard (1981) calls attention to the distinction between having each unit in a network represent a particular point in a parameter space (with its activation representing confidence in the validity of that point) and having units whose activation represents the value of a (necessarily one-dimensional) parameter. The former is called a *value unit* encoding by Ballard and is used extensively in his generalized Hough transform approach to early visual processing; the latter is called a *variable unit* encoding.

For purposes of this paper, call any variable unit encoding a *Type I* encoding; the class of *Type II* encodings will include any value unit encoding as well as any representation typified by pixel-level descriptions of retinal images.

As a concrete example, consider a 1-dimensional array of 10 units such that the only patterns which appear in this array all consist of two adjacent 1's with the rest 0's. This is a Type II encoding of a pattern space which may be considered essentially one-dimensional; the 10-dimensional vectors

which represent the patterns jump around in the space in such a way that this one-dimensionality is not easily recognized. This one-dimensionality is really a consequence of the manner in which the patterns overlap.

In contrast, this same pattern space may be given a Type I encoding in a single unit whose activation is a monotonic function of, say, the distance of the leftmost 1 in the pattern from the left-hand end of the array.

At this point, the central thesis of this paper can be stated: *Self-organizing mappings from Type I representations is straightforward; self-organizing mappings from Type II representations, if possible at all, will require the use of mechanisms yet to be discovered.* In support of the first half of this thesis, I present the following two examples of self-organizing mappings, the first taken from work of Kohonen (1982) and the second from recent work of my own. Following these examples is a discussion in support of the second half of this thesis.

Example 1. Let the A_2 layer have a certain pattern of lateral feedback connections so that the only patterns of activity which it supports are such that all non-zero activity is confined to a very small number of nearby units. In particular, assume that the units are laid out in 2-dimensional space in such a way that nearby units excite one another but more distant units inhibit one another. Suppose that the A_1 layer consists of 2 units, with patterns drawn uniformly from a convex subset of Euclidean 2-space. Suppose also that there are no $A_2 \rightarrow A_1$ connections. Then Kohonen (1982) has shown that, by using a common variant of what has come to be known as the Hebb learning rule, the $A_1 \rightarrow A_2$ mapping will generally self-organize in such a way that nearby units respond most strongly to nearby patterns. [1] The resulting mapping re-codes the 2-dimensional pattern space implicit in the activations of the A_1 units in such a way that its 2-dimensionality becomes explicit in the A_2 layer. In the language of Ballard (1981), the resulting mapping can be said to turn a variable unit encoding into what is essentially a value unit encoding; in the terminology of this paper, the resulting mapping recasts a Type I representation into a particular Type II representation.

Example 2. Let the system have no lateral connections in either the A_1 layer or the A_2 layer, but let there be reciprocal $A_2 \rightarrow A_1$ connections. Let the A_1 units apply a weight modification rule to their incoming $A_2 \rightarrow A_1$ connections which has the effect of trying to more closely match their current pattern; furthermore, let them apply this same correction to their outgoing $A_1 \rightarrow A_2$ connections. Then, if the bottom-up and top-down connections are symmetrical, [2] the system performs a principal component analysis of the training stimuli during self-organization. More precisely, let n_2 denote the number of units in the A_2 layer. Then, if this system is trained with patterns having mean 0, self-organization causes the output corresponding to any given input vector to consist of a projection onto the subspace spanned by the eigenvectors corresponding to the n_2 largest eigenvalues of the scatter matrix of the training stimuli. This output is expressed in some orthonormal basis which need not be these eigenvectors themselves. In other words, individual units in A_2 will not necessarily be feature detectors for individual principal components; instead the output encoding may be distributed with respect to these components. A fuller account of the details of this system and an analysis of its behavior will appear elsewhere.

---

1. I have slightly simplified the actual details of Kohonen's work in order to avoid discussion of technical matters not germane to this presentation.

2. These weights need not be assumed symmetrical at the outset; the simple trick of allowing all weights to decay slowly will accomplish the necessary symmetry eventually.

The key point to be made about the system of Example 2 in the context of this paper is that it readily self-organizes a useful mapping from one Type I representation to another.

While the work of Kohonen (1982) and Amari (1983) may leave one with the impression that certain Type II → Type II mappings may be self-organized in the same way as described in Example 1, I would claim that, in general, a good mapping is not achieved through the application of such a learning rule. For example, suppose that the A_1 layer is an identical copy of the A_2 layer as described in Example 1 and the system is expected to self-organize what is essentially an identity mapping. This is a simple version of the problem of forming a topographic map between, for example, the retina and visual cortex. While Amari (1983) makes certain claims about such self-organization being possible, he readily concedes that it is difficult to obtain a topographic map from such a system if one starts with totally random initial connections. In fact, my simulations of such a system would suggest that unless one starts with initial connections very close to what one intends as the final outcome, the system is very unlikely to form a true topographic map. The major difficulty, it appears to me, is that the learning rule basically requires that, at statistical equilibrium, the stimulus vector to which each A_2 unit most strongly responds must be equal to a weighted average of the stimulus vectors to which its neighbors (in the topology of the lateral connectivity of A_2) most strongly respond. This is the underlying reason why such a system works well for self-organizing convex Type I input, and why I claim that it cannot be expected to do the same for Type II input. Indeed, this fundamental difference is the main motivation behind my drawing the distinction between these two types of representation. This argument in fact suggests that any learning rule which causes an A_2 unit to learn to respond to an average of the stimulus vectors which have excited it cannot be expected to achieve a good mapping between Type II encodings. Some other learning rule must be used to achieve this.

Thus I argue that it remains an open question whether mappings can, in general, be self-organized from a Type II representation to another Type II representation. Discovery of such a mechanism would be quite interesting, since it is possible to specify lateral connectivity patterns in the A_2 layer which could force any particular topology on the stimulus space. As an example, if the A_2 layer has the connectivity of a Möbius band and the A_1 layer is a patch of 2-dimensional retina upon which patterns of activity are elongated bars of various orientation and position, then application of this mechanism should lead to a mapping in which each A_2 unit is maximally responsive to a particular combination of orientation and position. One can imagine self-organizing just about any Hough-style transform in this manner.

Another intriguing possibility which is suggested by this work is that of self-organizing a mapping from a Type II representation to a Type I representation. As an example of such a mapping, consider a description of a connected pattern on a 2-dimensional retina in terms of Fourier descriptors for the boundary (Zahn & Roskies, 1972; Persoon & Fu, 1977) along with the coordinates of the center of mass, all encoded in variable units. What is appealing about this particular example is that it should be much more economical in both units and connections to compute a mapping between a retina-based description of an object and an object-based description of that object (Hinton, 1981) if these descriptions are encoded in variable units than if they are encoded in value units.

References

Amari, S. (1983). Field theory of self-organizing neural nets. *IEEE Transactions on Systems, Man, and Cybernetics, 13*, 741-748.

Ballard, D. H. (1981). Parameter networks: Towards a theory of low-level vision. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, Vancouver, B.C., Canada, 1068-1078.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience, 2*, 32-48.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 6*, 205-254.

Grossberg, S. (1980). How does the brain build a cognitive code? *Psychological Review, 87*, 1-51.

Hinton, G. E. (1981). A parallel computation that assigns canonical object-based frames of reference. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, Vancouver, B.C., Canada, 683-685.

Hinton, G. E., & Sejnowski, T. J. (1983). Analyzing Cooperative Computation. *Proceedings of the Fifth Annual Conference of the Cognitive Science Society*, Rochester, NY, 683-685.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 49-69.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception, Part 1: An account of the basic findings. *Psychological Review, 88*, 375-407.

Nass, M. M., & Cooper, L. N. (1975). A theory for the development of feature detecting cells in the visual cortex. *Biological Cybernetics, 19*, 1-18.

Persoon, E., & Fu, K. (1977). Shape discrimination using Fourier descriptors. *IEEE Transactions on Systems, Man, and Cybernetics, 7*, 170-179.

Rosenblatt, F. (1961). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Washington, DC: Spartan.

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik, 14*, 85-100.

Zahn, C., & Roskies, R. (1972). Fourier descriptors for plane closed curves. *IEEE Transactions on Computers, 21*, 269-281.