Proceedings of the Int. Conf. on Parallel
Processing in Neural Sustems and Computers
(ICNC), Düsseldorf, (Elsevier, 1990) (in press)

A SELF-ORGANIZING NETWORK FOR COMPLETE FEATURE EXTRACTION

Jeanne RUBNER, Klaus SCHULTEN* and Paul TAVAN

Physik-Department
Technische Universität München
8046 Garching, Federal Republic of Germany

We describe a two-layered network of linear neurons that organizes itself as to extract the maximal amount of information contained in a set of presented patterns. The weights between layers obey a Hebbian rule, whereas the lateral, hierarchically organized weights within the output layer follow an Anti-Hebbian rule. For a proper choice of the learning parameters, this rule forces the activities of the output units to become uncorrelated and the lateral weights to vanish. The weights between the two layers converge to the eigenvectors of the covariance matrix of input patterns, i.e. the network performs a principal component analysis of the input information. Consequently the output units become detectors of orthogonal features, similar to ones found in the brain of mammals.

1. INTRODUCTION

Although part of the synaptic connections in the brain is genetically specified, postnatal visual input plays an essential role in the organization, birth and death of synapses. One expects that local rules, like Hebb's rule [1], govern the postnatal organization of the brain and the formation of feature detectors or visual filters. These expectations raise the general question how a sensory system, in response to input information, can organize itself according to local rules so as to form feature detectors which encode mutually independent aspects of the information contained in patterns presented to it.

In the following section we describe a simple, two-layered neural network as a model for such a system. We sketch the mathematical properties of this model and present results of simulations yielding visual filters that respond to patterns of varying orientations and spatial frequencies.

2. THE NETWORK MODEL

The network consists of an input and an output layer with N_i and N_o neurons, respectively. The units exhibit real, continuous-valued activities $\mathbf{i} = (i_1, .., i_{N_i})$ and $\mathbf{o} = (o_1, .., o_{N_o})$. The two layers are completely interconnected, and the weight of the connection between input unit j and output unit m is denoted by w_{jm}. The set of weights connecting an output unit m to all input units forms the weight vector $\mathbf{w}_m$, the transpose of which is the m-th row of the weight matrix $\mathbf{W}$. The set of N_π presented patterns is denoted by $\{\mathbf{p}^\pi = (p_1^\pi, .., p_{N_i}^\pi), \pi = 1, \ldots, N_\pi\}$. The activities of the input units correspond to the presented patterns, i.e., $\mathbf{i} = \mathbf{p}^\pi$, the activities of the output units are the sums of the inputs weighted by the synaptic strengths, i.e., $\mathbf{o}^\pi = \mathbf{W}\,\mathbf{p}^\pi$.

The weights between the two layers are adjusted upon presentation of an input pattern $\mathbf{p}^\pi$ according to a Hebbian rule, i.e., $\Delta \mathbf{w}_m = \eta\, \mathbf{p}^\pi o_m^\pi$ with positive η [1]. As suggested in Ref. [4], we choose the pattern set such that $\langle \mathbf{p}^\pi \rangle = 0$. Here, the brackets $\langle \ldots \rangle$ denote the average over the set of patterns.

If the network contains a single output unit, the Hebbian rule and an Euclidean normalization of weights after every update, i.e., $\sum_i w_{i1}^2 = 1$, render weights which characterize the direction of maximal variance of the pattern set [4], [5]. Equivalently, the weightvector $\mathbf{w}_1$ converges

*present adress: Department of Physics, University of Illinois, Urbana, Ill 61801, USA to the eigenvector with the largest eigenvalue of the covariance matrix $\mathbf{C}$ of the pattern set, whose elements are given by are given by $C_{jk} = \langle p_j^\pi p_k^\pi \rangle$. Diagonalizing a covariance matrix corresponds to the statistical technique of principal component analysis [6]. Thus, a Hebbian learning rule for normalized weights yields the first principal component of the input data set. Consequently, the output unit corresponds to a visual filter extracting the most important feature contained in the set of presented patterns.

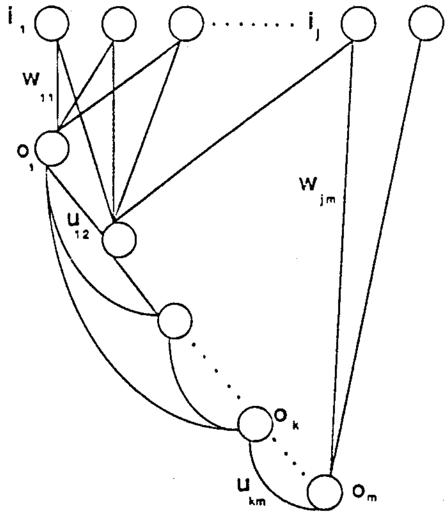

Fig.1 Scheme of the proposed network.

However, a single unit only processes a fraction of the total information contained in a pattern. In order to transmit the complete information between the two layers, more output cells and a different learning rule are required. As a result of an unsupervised learning process, the weight vectors of these units should represent the remaining principal components, i.e., the remaining eigenvectors of $\mathbf{C}$. For that purpose we assume the existence of lateral, hierarchically organized connections with weights u_{lm} between the output units. Then the activity of the m-th output cell is given by $o_m^\pi = \mathbf{w}_m \cdot \mathbf{p}^\pi + \sum_{l<m} u_{lm} \mathbf{w}_l \cdot \mathbf{p}^\pi$. Figure 1 shows a scheme of the network. We propose that these lateral weights adapt themselves according to an *anti-Hebbian* rule: the change of the lateral synaptic weight u_{lm} between two output units l and m is negatively proportional to the product of pre- and postsynaptic activities, namely

$$\Delta u_{lm} = -\mu \, o_l^\pi \, o_m^\pi, \tag{1}$$

where μ is a positive learning parameter. This strictly local rule forces the lateral weights to vanish and the activities of the output cells to become uncorrelated. Correspondingly, the weight vectors $\mathbf{w}_m$ converge to the eigenvectors $\mathbf{c}_\alpha$ of the covariance matrix $\mathbf{C}$. As a result, the output units correspond to analyzers of mutually orthogonal features that extract the directions of diminishing variance of the input patterns.

3. MATHEMATICAL ANALYSIS

Consider first the case of N_i input and $N_o = 2$ output cells, connected by the lateral weight $u \equiv u_{12}$. For slow weight changes, average products of pre- and postsynaptic activities can be used in the learning rules. Then it is possible to expand the weight vectors in terms of eigenvectors $\mathbf{c}_\alpha$ of $\mathbf{C}$ and to derive differential equations for the expansion coefficients $d_{1\alpha}$, $d_{2\alpha}$ and for the lateral connection u, namely

$$\dot{d}_{m\alpha} = -d_{m\alpha} + \frac{(1+\eta\lambda_\alpha)d_{m\alpha} \; + \; \delta_{m2}\,\eta\, u\, \lambda_\alpha\, d_{1\alpha}}{\sqrt{\sum_\beta \left[(1+\eta\lambda_\beta)d_{m\beta} \; + \; \delta_{m2}\,\eta\, u\, \lambda_\beta\, d_{1\beta}\right]^2}} \equiv f_{d_{m\alpha}}(d_{m\beta}, u), \tag{2}$$

and

$$\dot{u} = -\mu \sum_{\beta} \lambda_\beta d_{1\beta} d_{2\beta} - \mu u \sum_{\beta} \lambda_\beta d_{1\beta}^2 \equiv f_u(d_{m\beta}, u), \tag{3}$$

where the dot denotes the time derivative and λ_β the β-th eigenvalue of $\mathbf{C}$ ($\lambda_\beta \geq \lambda_{\beta+1}$). Carrying out a linear stability analysis for this system of coupled differential equations [3], one finds upper and lower limits for the learning parameter μ,

$$\frac{2}{\lambda_1} > \mu > z^{(2)} = \frac{\eta \Delta\lambda}{\lambda_1(1+\eta\lambda_2)}. \tag{4}$$

In order to get analytical results for the case of more than two output units, we assume that $n-1$ of the total N_o weight vectors have already converged to the first $n-1$ eigenvectors of $\mathbf{C}$ and that the lateral weights between the corresponding output units have vanished. (Note that a corresponding procedure is not necessary in practice, since our simulations have shown that the weight vectors converge simultaneously). Then, for reasons of symmetry, only those variables u_{mn} and $d_{n\alpha}$ are coupled for which $m = \alpha < n$. Requiring a stable fixpoint for the corresponding differential equations, again yields a lower limit for μ, namely

$$\mu > z^{(n)} = \frac{\eta(\lambda_1 - \lambda_n)}{\lambda_1(1+\eta\lambda_n)}. \tag{5}$$

We have checked these analytical results by applying the proposed learning scheme to one-dimensional, random patterns with nearest-neighbor correlations. This corresponds to diagonalizing the tight-binding Hamiltonian of a linear chain of atoms, the eigenvalues and eigenvectors of which are well-known. Thus, we can compute $d_{m\alpha}(t)$ as well as the upper and lower limits for μ. The numerical behavior of the weight vectors and lateral weights is in excellent agreement with the analytically predicted behavior [3].

4. ORIENTATION AND SPATIAL FREQUENCY SELECTIVE CELLS

In the following, we examine which are the essential features of spatially varying patterns and compare the receptive fields obtained by our learning scheme with the ones of simple cells, feature detectors selective to edges or bars, which represent the first stage of spatial information processing in the primary visual cortex [7].

For this purpose, we consider a rectangular lattice of $N_i \times N_i'$ sensory input units representing the receptive field of N_o output units, with $N_o \leq N_i N_i'$. We generate two-dimensional patterns of varying intensity by first selecting random numbers from the interval $[-1, +1]$. Then, in order to introduce information about the topological structure of the receptive field, the random input intensities are correlated, e.g., with their nearest neighbors in both directions. We assume vanishing boundary conditions. Note, that this averaging of neighboring signals corresponds to introducing an additional layer with random activities and with fixed and restricted connections to the input layer.

Receptive fields of simple cells in cat striate cortex can be described by Gabor functions [8], which consist of an oscillatory part, namely a sinusoidal plane wave and a Gaussian, exponentially decaying part. In analogy to the one-dimensional case of random patterns, one expects receptive fields corresponding to the eigenfunctions of a tight binding lattice of atoms, i.e., sinusoidal plane waves with vanishing boundary conditions. In order to implement an exponential decay, we scale the weights between layers by a two-dimensional Gaussian distribution centered at the lattice location $(N_i/2, N_i'/2)$ with widths in x- and y-directions σ_1 and σ_2. The non-homogeneous distribution of weights between layers could correspond to a higher density of nearby input cells.

If the Gaussian distribution is not rotationnally symmetric (i.e., if $\sigma_1/\sigma_2 \neq 1$), degeneracy of eigenvalues is broken and mixing of eigenfunctions does not occur. However, the orientation of

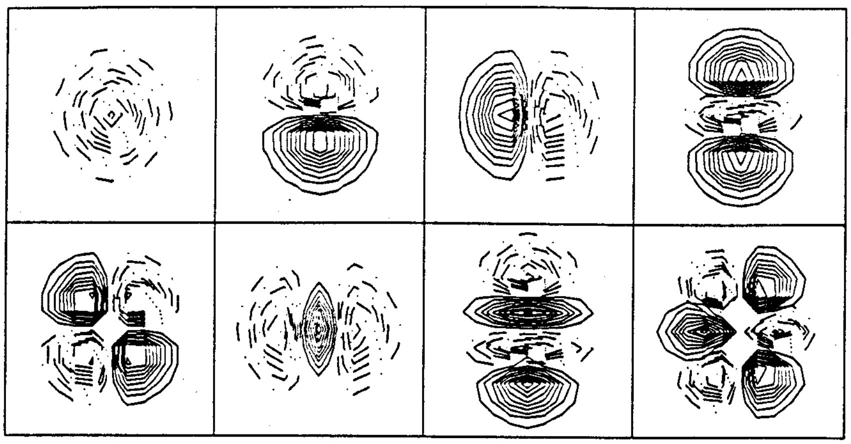

Fig.2 From left to right and top to bottom: contour plots of receptive fields of output units 1-8 in the case of a square lattice of 20 × 20 input units.

receptive fields is pre–determined, due to imposed symmetry axes. Figure 2 displays contour plots of the receptive fields of the first eight output cells after 10000 learning cycles (from left to right and top to bottom). Solid lines correspond to positive, dashed lines to negative synaptic weights. The input lattice was square, with 20 × 20 units and the parameters of the Gaussian distribution of weights were $\sigma_1 = 11$ and $\sigma_2 = 14$. Learning parameters η and μ were equal to 0.05 and 0.1, respectively. Due to the non symmetric Gaussian distribution of weights, all units have slightly elongated receptive fields. The first unit corresponds to a simple cell with all-inhibitory synaptic weights. The receptive fields of the second and third units display an excitatory and an inhibitory region and resemble simple cells, selective to edges of a fixed orientation. The fourth and sixth units have receptive fields with two zero-crossings, corresponding to simple cells, selective to bars of a fixed orientation. The seventh unit is as well orientation selective, with four alternating excitatory and inhibitory regions. This unit would maximally respond to two parallel lines or bars with fixed distance and orientation. All the described units have receptive fields that resemble recorded receptive fields of simple cells in the primary visual cortex [8]. Up to now, there has not been any experimental evidence for receptive fields of the type of the fifth and eighth units, displaying four and six lobes. However, if the scheme of spatial information processing in terms of a local Fourier analysis is correct, such receptive fields might exist in the visual cortex.

ACKNOWLEDGEMENTS

This work has been supported by the Bundesministerium für Forschung und Technologie (ITR-8800-G9) and the Deutsche Forschungsgemeinschaft SFB 143-C1.

REFERENCES

[1] Hebb, D.O., The Organization of Behavior (Wiley, New York, 1949).
[2] Rubner, J. and Schulten, K., Biol Cybern, in press.
[3] Rubner, J. and Tavan, P., Europhys Lett, in press.
[4] Oja, E., J Math Biology 15 (1982) 267.
[5] Linsker, R., IEEE Computer March (1988) 105.
[6] Lawley, D.N. and Maxwell. A.E., Factor Analysis as a Statistical Method (Butterworths, London, 1963).
[7] Hubel, D.H. and Wiesel, T.N., J Physiol 160 (1962) 106.
[8] Jones, J.P. and Palmer, L.A., J Neurophysiol 58 (1987) 1187.