In: Proceedings of the International Conference on Artificial Neural Networks 2001

Self-Organization of Orientation Maps, Lateral Connections, and Dynamic Receptive Fields in the Primary Visual Cortex

C. Weber

Dept. of Brain and Cognitive Science, University of Rochester, NY, USA

Abstract. We set up a combined model of sparse coding bottom-up feature detectors and a subsequent attractor with horizontal weights. It is trained with filtered grey-scale natural images. We find the following results on the connectivity: *(i)* the bottom-up connections establish a topographic map where orientation and frequency are represented in an ordered fashion, but phase randomly. *(ii)* the lateral connections display local excitation and surround inhibition in the feature spaces of position, orientation and frequency. The results on the attractor activations after an interrupted relaxation of the attractor cells as a response to a stimulus are: *(i)* Contrast-response curves measured as responses to sine gratings increase sharply at low contrasts, but decrease at higher contrasts (as reported for cells which are adapted to low contrasts [1]). *(ii)* Orientation tuning curves of the attractor cells are more peaked than those of the feature cells. They have reasonable contrast invariant tuning widths, however, the regime of gain (along the contrast axis) is small before saturation is reached. *(iii)* The optimal response is roughly phase invariant, if the attractor is trained to predict its input when images move slightly.

Introduction

It is common for hierarchical maximum likelihood models to assume a factorial prior probability distribution of the hidden variables on each layer. However, statistical dependencies remain to be captured by subsequent hierarchical levels. Lateral connections can help to find this structure, may they serve to statistically de-correlate [10] or to find correlation structure [2] within activities of nearby cells. Lateral connections which form an attractor were shown [5] to recover noisy input with maximum likelihood, given certain noise models. Interpreting shifts in images as noise, may supply an approximate theoretical framework for shift invariant recognition as well as for understanding complex cells in V1.

In our contribution, bottom-up weights W^{bu} of hidden layer neurons ("feature cells") and the top-down feed-back projections V^{td} (Fig. 1) are trained with a sparse coding paradigm to become topographically arranged edge detectors. We train lateral connections V^{lat} to form attractors which capture the correlation structure on the activations of hidden layer neurons. The "attractor cells" receive initial input from the feature cells and then relax freely for a fixed period of time using the lateral connections. The final activation state is defined as their

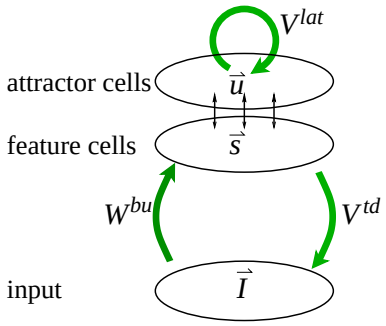


Fig. 1. Model architecture with weights W^{bu}, V^{td}, V^{lat} and activations $\boldsymbol{I}$, $\boldsymbol{s}$, $\boldsymbol{u}$.

response to the input and compared to biological cell response properties in V1. We find that this simple setup can explain a variety of biological observations.

Theory and Methods

The Helmholtz machine framework [3] is originally intended as a purely hierarchical model: generative, top-down weights are trained to generate the data while bottom-up, recognition weights are trained to statistically invert the generative model. Hereby, the data can be considered as part of the recognition model and supply the target values for training the generative model. Vice versa, the prior on the hidden layer activations is considered as part of the generative model and supplies the target training values for the recognition model.

However, the recognition model should be optimized using activation statistics as delivered by the training data instead of the prior on the hidden layer activations [3]. This can be done by lateral weights [2]: they should learn the statistics of the hidden unit activations during presentation of the data in order to recall these data during the *sleep phase* to be used as target values for training the recognition model. Lateral weights are thus part of the generative model.

Bottom-up and top-down weights However, to train the lateral weights, a coarsely functioning recognition model has to exist first. For clarity, we do not use the prior which is learned by the attractor cells for training of the recognition weights at all. Instead, we assume a topographic Gaussian function as a prior on "higher order neurons" [7] to establish a topographically ordered orientation map.

Recently, the square of feature cell responses has been proposed as the relevant feature to arrange neurons topographically on V1 [8][7], because this will produce a random phase distribution as observed experimentally [4]. To mimic this effect, we do not distinguish excited from inhibited cells: in the *sleep phase*, we draw, from the topographic Gaussian prior, values (0, 1, 1′) on the feature cells as the target training values $\boldsymbol{s}^-$ for the recognition model. Hereby, a small number of mostly adjacent units is switched ON. The activity 1′ is interpreted

as an activation of an "inhibited" cell, and is thus used as a target value while the bottom-up input is inverted (by setting β, Eq. 5, to $-\beta$). The bottom-up and top-down connection matrices W^{bu}, V^{td} are used to transform the input $\boldsymbol{I}$ into a hidden representation $\boldsymbol{s}$, and vice versa:

$$I_j \;=\; \textstyle\sum_i v_{ji}^{td} s_i \;, \qquad\qquad s_i \;=\; f(\textstyle\sum_j w_{ij}^{bu} I_j) \tag{1}$$

They are trained to functionally invert each other mutually, by maximizing the log-likelihood to reconstruct a stimulus sent through the other weight matrix:

$$\Delta w_{ij}^{bu} \;\approx\; (s_i^- \;-\; \tilde{s}_i^-)\, I_j^- \;, \qquad\qquad \Delta v_{ji}^{td} \;\approx\; (I_j^+ \;-\; \tilde{I}_j^+)\, s_i^+ \tag{2}$$

where the upper index "$-$" means that a random hidden vector $\boldsymbol{s}^-$ initiated the activities (*sleep phase*) and "$+$" denotes initiation by data $\boldsymbol{I}^+$ (*wake phase*).

Lateral weights In the *wake phase*, attractor cell activations $\tilde{u}^+(t=0)$ are initialized with the maximum of the (positive) values $\boldsymbol{s}^+$ of an excitatory and its associated inhibitory neuron (this could as well be the sum instead of the maximum, because the smaller of the two values is almost vanishing). Then, the lateral weights V^{lat} define their activation dynamics through time:

$$\tilde{u}_i(t+1) \;=\; f(\textstyle\sum_l v_{il}^{lat} \tilde{u}_l(t)) \tag{3}$$

They are trained in the *wake phase* only, to generate (predict) incoming activities. Learning maximizes the log-likelihood to generate this distribution via Eq. 3:

$$\Delta v_{il}^{lat} \;\approx\; \textstyle\sum_t (u_i^+(t) \;-\; \tilde{u}_i^+(t))\, \tilde{u}_l^+(t-1)\,. \tag{4}$$

During each relaxation, input images were shifted by one pixel into a random direction at each time step. Hereby, "target" data $u_i^+(t)$ moved slightly away from $u_i^+(t=0) = \tilde{u}^+(t=0)$ which also initiated the relaxation process.

Parameters and sparseness Model neurons are binary with activations $\{0,1\}$ and probabilistic. We use the probability for neuron i to be active as a continuous transfer function:

$$f(h_i) \;=\; p_i(1) \;=\; \frac{e^{\beta h_i}}{e^{\beta h_i} + n} \tag{5}$$

Parameters $\beta = 2$ scales the slope of the function and n is the degeneracy of the 0-state. Large $n = 64$ for the feature cells reduces the probability of the 1-state favoring sparseness. $n = 8$ for the attractor cells. A weight decay of $-0.03 \cdot w_{ij}$ on W^{bu} and V^{td} adds to Eqs. 2 to hinder weights to compensate for sparseness. W^{bu} and V^{td} were initialized randomly under topographic Gaussian envelopes, V^{lat} randomly, and self-connections were constantly set $v_{ii}^{lat} = 0$. Images were filtered as described in [9]. Input layer size: 16×16, hidden layer size: 24×24.

Results

Anatomy The hidden neurons have developed localized, Gabor function shaped receptive fields to the input. Fig. 2 shows essential observable parameters which

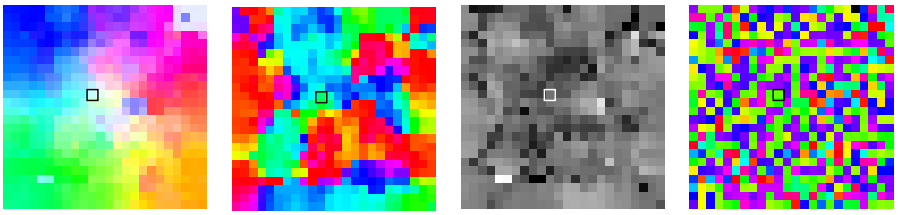


Fig. 2. Individual parameter maps for the input receptive fields of the simple hidden layer cells, from W_{10}. Parameters, except for spatial phase, vary continuously across large regions of the neuronal layer. A box marks the cell at position α (see Figs. 4,5).

were derived from Gabor function fits to the bottom-up connections W_{10}. Position, orientation and spatial frequency of the receptive fields in the input show continuous changes across the hidden layer, spatial phase w.r.t the receptive field center is distributed randomly.

Functional lateral connection densities along individual parameter axes are displayed in Fig. 3. The concept of local excitation and surround inhibition is well attributed to position-, orientation- and spatial frequency difference. Inhibition is for example strongest at cross-orientation, $\pi/2$. Lateral connections integrate over larger parts of phase space.

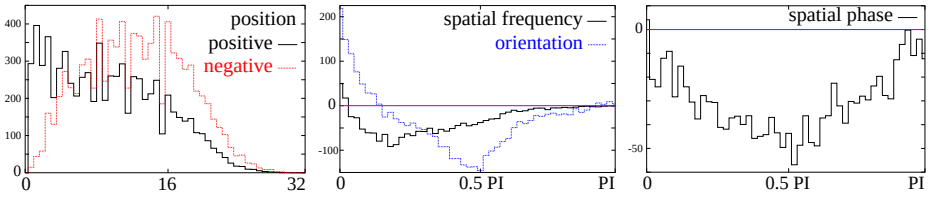


Fig. 3. Densities of lateral connections V^{lat}. **Left:** excitatory and inhibitory connections over distance across the hidden layer. **Middle, right:** functional connections (sum of excitatory and inhibitory) over feature differences. On the x-axes, differences between the features for two feature cell pairs (obtained from the Gabor fits to W_{10}) are displayed, binned into 50 intervals. The y-axis shows the sum of lateral connections which link attractor cell pairs whose feature difference is in the corresponding interval.

Physiology Fig. 4 shows the relaxation of the network activities after initialization with a sinusoidal grating. For learning and for obtaining the response curves below, relaxation was performed until time step 10. The response remains stable for limited further time.

Fig. 5 a) shows contrast response curves of *(i)* a feature cell and *(ii)* the corresponding attractor cell after relaxation ($t = 10$). With increasing contrast

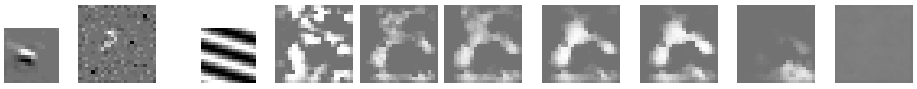
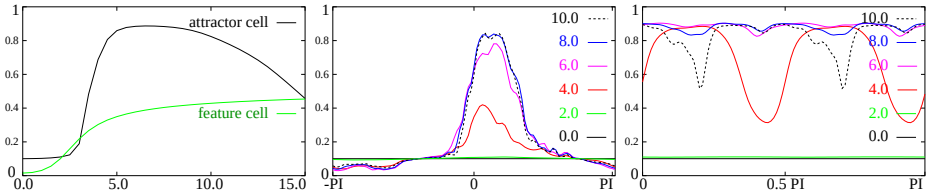


Fig. 4. From left to right: the receptive field $\boldsymbol{w}_\alpha^{BU}$ of feature cell indexed α. The lateral weights $\boldsymbol{v}_\alpha^{lat}$ of the corresponding attractor cell. Bright positive, dark negative values. A sinusoidal vector $\boldsymbol{I}$ shown at the **input**. Starting from $t = 0$, attractor activations $\boldsymbol{u}(t)$ develop after initialization from that stimulus through times given later on.

(simulated as stimulus amplitude), the feature cell's response follows its transfer function. In contrary, the attractor cell's response within the lateral dynamics saturates, however, unlike biologically observed, near its "physically" maximal activation. It decreases at higher contrasts, correctly [1]. Note that a sine grating activates more neurons than a typical image as used for training.

Orientation tuning curves of the attractor cell, Fig. 5 b), display a more or less constant and sharp tuning width. At cross-orientations and high contrasts, there is suppression w.r.t. zero-input activity.

Fig. 5 **c)** shows that the attractor neuron responds across a large phase range, with phase dependence only in a small contrast range. To obtain phase invariance, the "target" data $u_i^+(t)$ for the attractor must not be constant in time. Biologically, micro-saccades may deliver slightly shifting input.



a) contrast **b)** orientation **c)** spatial phase

Fig. 5. a) Contrast response curves, for the attractor cell at position α and the corresponding feature cell. The average over phase is taken, which downsizes the output of the feature cell. **b)** Orientation tuning and **c)** phase tuning curves of the attractor cell at different contrasts. Input was a static sine wave pattern with the optimal parameters for the corresponding feature cell (which in **b)** defines the 0-orientation as optimal).

Discussion

Single new elements in this work are only a Helmholtz machine which develops edge detectors from images or forms a topography. However, to our knowledge, a network fully trained on natural images and (complex) cells with contrast invariant orientation tuning are for the first time integrated.

The present work investigates the learning mechanisms which produce biologically observed results in this system, assuming that the architectural setting is suited. We show that learning the correlation structure in the activations which are initiated by the data (Eq. 4) produces weights with a structure of local excitation and surround inhibition. This is sufficient to generate sharp, contrast invariant orientation tuning curves. In particular, we do **not** need: *(i)* a sparseness constraint on the attractor activities to sharpen the orientation tuning curves, or *(ii)* de-correlation [6] of the attractor unit activations.

Fortunately, the attractor cannot generate its input data correctly (even if images were not shifted and the identity was learned). Otherwise, broadly tuned, feature cell response properties would remain. A (line-) attractor is instead the only patterns that the attractor cells can maintain. Each attractor is a hill of activity on similar neurons, but more irregular than if constructed analytically [11].

The fact that we observe different orientation tuning curves at different contrasts reflects a dynamic effect of the relaxation procedure. At low contrasts the hill of activity builds up from smaller activations, primarily involving specific excitatory connections to the few active cells. At high contrasts, suppression via less specific but more numerous inhibitory connections dominates.

In our model set-up, the input to the attractor is not controlled by parameters and does not constantly influence the attractor. The question remains, how the attractor cells can *learn* to select their input from the feature cells.

References

1. M. Carandini and D. Ferster. Membrane potential and firing rate in cat primary visual cortex. *J. Neurosci.*, 20(1):470–84, 2000.
2. P. Dayan. Recurrent sampling models for the Helmholtz machine. *Neur. Comp.*, 11:653–77, 2000.
3. P. Dayan, G. E. Hinton, R. Neal, and R. S. Zemel. The Helmholtz machine. *Neur. Comp.*, 7:1022–1037, 1995.
4. G.C. DeAngelis, G.M. Ghose, I. Ohzawa, and R.D. Freeman. Functional micro-organization of primary visual cortex: Receptive field analysis of nearby neurons. *J. Neurosci.*, 19(9):4046–64, 2000.
5. S. Deneve, P.E. Latham, and A. Pouget. Reading population codes: a neural implementation of ideal observers. *Nature Neurosci.*, 2(8):740–5, 1999.
6. D.W. Dong. Associative decorrelation dynamics: A theory of self-organization and optimization in feedback networks. In *Proceedings of NIPS 7*, pages 925–32, 1994.
7. A. Hyvärinen and P.O. Hoyer. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neur. Comp.*, 12:1705–20, 2000.
8. N. Mayer, J.M. Herrmann, H.U. Bauer, and T. Geisel. A cortical interpretation of ASSOMs. In *Proceedings ICANN*, pages 961–966, 1998.
9. B.A. Olshausen and D.J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.
10. J. Sirosh and R. Miikkulainen. Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. *Neur. Comp.*, 9:577–94, 1997.
11. K. Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *J. Neurosci.*, 16:2112–26, 1996.