Communicated by Geoffrey Hinton

Learning Invariance from Transformation Sequences

Peter Földiák
Physiological Laboratory, University of Cambridge,
Downing Street, Cambridge CB2 3EG, U.K.

The visual system can reliably identify objects even when the retinal image is transformed considerably by commonly occurring changes in the environment. A local learning rule is proposed, which allows a network to learn to generalize across such transformations. During the learning phase, the network is exposed to temporal sequences of patterns undergoing the transformation. An application of the algorithm is presented in which the network learns invariance to shift in retinal position. Such a principle may be involved in the development of the characteristic shift invariance property of complex cells in the primary visual cortex, and also in the development of more complicated invariance properties of neurons in higher visual areas.

1 Introduction

How can we consistently recognize objects when changes in the viewing angle, eye position, distance, size, orientation, relative position, or deformations of the object itself (e.g., of a newspaper or a gymnast) can change their retinal projections so significantly? The visual system must contain knowledge about such transformations in order to be able to generalize correctly. Part of this knowledge is probably determined genetically, but it is also likely that the visual system learns from its sensory experience, which contains plenty of examples of such transformations. Electrophysiological experiments suggest that the invariance properties of perception may be due to the receptive field characteristics of individual cells in the visual system. Complex cells in the primary visual cortex exhibit approximate invariance to position within a limited range (Hubel and Wiesel 1962), while cells in higher visual areas in the temporal cortex show more complex forms of invariance to rotation, color, size, and distance, and they also have much larger receptive fields (Gross and Mishkin 1977, Perrett *et al.* 1982). The simplest model of a neuron, which takes a weighted sum of its inputs, shows a form of generalization in which patterns that differ on only a small number of input lines generate similar outputs. For such a unit, patterns are similar when they are close in Hamming distance. Any simple transformation, like a shift in position or a rotation, can cause a great difference in Hamming distance, so this

simple unit tends to respond to the transformed image very differently and generalizes poorly across the transformation. The solution to this problem is therefore likely to require either a more complex model of a neuron, or a network of simple units.

2 Shift Invariance

Fukushima (1980) proposed a solution to the positional invariance problem by a network consisting of alternating feature detector ("S" or simple) and invariance ("C" or complex) layers. Feature detectors in the "S" layer are replicated in many different positions, while the outputs of detectors of the same feature are pooled from different positions in the "C" layers. The presence of the feature in any position within a limited region can therefore activate the appropriate "C" unit. This idea is consistent with models of complex cells in the primary visual cortex (Hubel and Wiesel 1962; Spitzer and Hochstein 1985) in that they assume that complex cells receive their major inputs from simple cells or simple-cell-like subunits selective for the same orientation in different positions. In Fukushima's model, the pair of feature detecting and invariance layers is repeated in a hierarchical way, gradually giving rise to more selectivity and a larger range of positional invariance. In the top layer, units are completely indifferent to the position of the pattern, while they are still sensitive to the approximate relative position of its components. In this way, not only shift invariance, but some degree of distortion tolerance is achieved as well. This architecture has successfully been applied both by Fukushima (1980) and LeCun *et al.* (1989) in pattern recognition problems. LeCun *et al.* achieve reliable recognition of handwritten digits (zip codes) by using such architectural constraints to reduce the number of free parameters that need to be adjusted. Some of the principles presented in these networks may also be extremely helpful in modeling the visual system. The implementation of some of their essential assumptions in biological neural networks, however, seems very difficult. Apart from the question of the biological plausibility of the backpropagation algorithm used in LeCun *et al.*'s model, both models assume that the feature detectors are connected to "complex" units in a fixed way, and that all the simple units that are connected to a complex unit have the same input weight vector (except for a shift in position). Therefore whenever the weights of one of the "simple" units are modified (e.g., by a Hebbian mechanism), the corresponding weights of all the other simple units connected to the same complex unit need to be modified in exactly the same way ("weight sharing"). This operation is nonlocal for the synapses of all the units except for the one that was originally modified. A "learn now" signal broadcast by the complex unit to all its simple units would not solve this problem either, as the shifted version of the input, which would be necessary for local learning, is not available for the simple units.

3 A Learning Rule

An arrangement is needed in which detectors of the same feature all connect to the same complex unit. However, instead of requiring simple units permanently connected to a complex unit (a "family") to develop in an identical way, the same goal can be achieved by letting simple units develop independently and then allowing similar ones to connect adaptively to a complex unit (form "clubs"). A learning rule is therefore needed to specify these modifiable simple-to-complex connections. A simple Hebbian rule, which depends only on instantaneous activations, does not work here as it only detects overlapping patterns in the input and picks up correlations between input units. If the input to the simple layer contains an example of the feature at only one spatial position at any moment then there will never be significant overlap between detectors of that feature in different positions. The absence of positive correlations would prevent those units being connected to the same output. The solution proposed here is a modified Hebbian rule in which the modification of the synaptic strength at time step t is proportional not to the pre- and post-synaptic activity, but instead to the presynaptic activity (x) and to an average value, a trace of the postsynaptic activity ($\overline{y}$). A second, decay term is added in order to keep the weight vector bounded:

$$\Delta w_{ij}^{(t)} = \alpha \overline{y}_i^{(t)} (x_j^{(t)} - w_{ij}^{(t)})$$

where

$$\overline{y}_i^{(t)} = (1 - \delta)\overline{y}_i^{(t-1)} + \delta y_i^{(t)}$$

A similar trace mechanism has been proposed by Klopf (1982) and used in models of classical conditioning by Sutton and Barto (1981). A trace is a running average of the activation of the unit, which has the effect that activity at one moment will influence learning at a later moment. This temporal low-pass filtering of the activity embodies the assumption that the desired features are stable in the environment. As the trace depends on the activity of only one unit, the modified rule is still local. One possibility is that such a trace is implemented in a biological neuron by a chemical concentration that follows cell activity.

4 Simulation

The development of the connections between the simple and complex units is simulated in an example in which the goal is to learn shift invariance. In the simple layer there are position-dependent line detectors, one unit for each of 4 orientations in the 64 positions on an 8×8 grid. There are only 4 units in the complex layer, fully connected to the simple units. During training, moving lines selected at random from four orientations and two directions are swept across a model retina, which gives

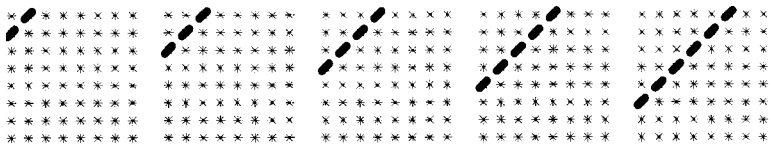

Figure 1: Five consecutive frames from one of the sequences used as input. Each line segment represents a simple unit of the corresponding orientation and position. Thick segments are active ($x_j = 1$), thin ones are inactive units ($x_j = 0$). The trace is maintained between sweeps.

rise to activation of the simple units of the appropriate orientation in different positions at different moments in time (Fig. 1). Such activation can either be the result of eye movements, object motion in the environment, or it may even be present during early development as there is evidence for waves of activity in the developing mammalian retina (Meister *et al.* 1990). The activation of these simple units is the input to the network. If an active simple unit succeeds in exciting one of the four complex units, then the trace of that complex unit gets enhanced for a period of time comparable to the duration of the sweep across the receptive fields of the simple units. Therefore all the connections from the simple units that get activated during the rest of that sweep get strengthened according to the modified Hebb rule. Simple units of only one orientation get activated during a sweep, causing simple units of only one orientation to connect to any given complex unit. To prevent more than one complex unit from responding to the same orientation, some kind of competitive, inhibitory interaction is necessary between the complex units. In some previous simulations an adaptive competitive scheme, decorrelation, was used (Barlow and Földiák 1989; Földiák 1990), which is thought to be advantageous for other reasons. For the sake of clarity, however, the simplest possible competitive scheme (Rumelhart and Zipser 1985) was used in the simulation described here. Each unit took a sum of its inputs weighted by the connection strengths. The output y_k of the unit with the maximal weighted sum was set to 1, while the outputs of the rest of the units were set to 0:

$$y_k = \begin{cases} 1 & \text{if } \mathrm{argmax}_i(\sum_j w_{ij}x_j) = k \\ 0 & \text{otherwise} \end{cases}$$

Figure 2a shows the initially random connections between the simple and the complex units, while Figure 2b shows the connections after training with 500 sweeps across the retina.

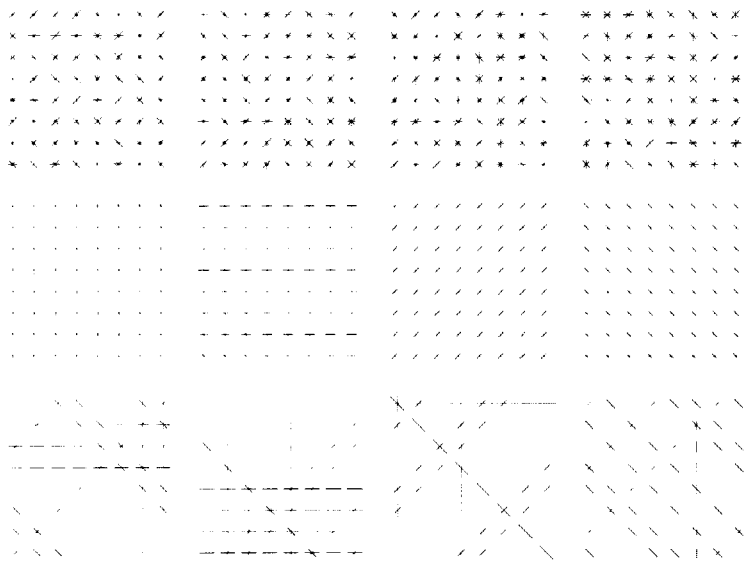

Figure 2: Connection patterns of the four complex units (a) before training and (b) after training on 500 line sweeps across the retina. The length of each segment indicates the strength of the connection from the simple unit of the corresponding position and orientation to the complex unit. Initial weights were chosen from a uniform distribution on $[0, 0.1]$. $\alpha = 0.02$, $\delta = 0.2$. (c) The result of training without trace ($\delta = 1$).

5 Discussion

The simple example given above is not intended to be a realistic model of complex cell development, since unoriented input to complex cells was ignored and simple units were considered merely as line detectors. By using a more realistic model of simple cells, the above principle would be able to predict that simple cells of the same spatial frequency and orientation but of different phase tuning (dark/bright line centre, even/odd symmetry) connect to the same complex cell, which would therefore lose sensitivity to phase. A further consequence would be that simple cells tuned to different spatial frequencies would segregate on different complex cells. The application of this algorithm to more complicated or abstract invariances (e.g., 3D rotations or deformations) would perhaps be even more interesting as it is even harder to see how they could be specified without some kind of learning; the way in which such invariance

properties could be wired in is much less obvious than in the case of positional invariance in Fukushima's or LeCun's networks. All that would be required by the proposed algorithm from previous stages of processing is that the transformation-dependent features should be available as input, and that the environment should generate sequences of the transformation causing the activation of these transformation-dependent detectors within a short period of time. Where no such detectors are available, other learning rules, based on temporal sequences or variation in form (Mitchison 1991, Webber 1991) may be able to find stable representations. If a supervision signal indicates the invariant properties, or self-supervision between successive time steps is applied, then backpropagation can also give rise to invariant feature detectors without explicit weight sharing (Hinton 1987). Nevertheless such learning is rather slow. Achieving a transformation-independent representation would certainly be very useful in recognizing patterns, yet the information that these invariance stages throw away may be vital in performing visual tasks. A "where" system would probably have to supplement and cooperate with such a "what" system in an intricate way.

Acknowledgments

I would like to thank Prof. H. B. Barlow, Prof. F. H. C. Crick, Dr. A. R. Gardner-Medwin, Prof. G. E. Hinton, and Dr. G. J. Mitchison for their useful comments. This work was supported by an Overseas Research Studentship, a research studentship from Churchill College, Cambridge, and SERC Grants GR/E43003 and GR/F34152.

References

Barlow, H. B., and Földiák, P. 1989. Adaptation and decorrelation in the cortex. In *The Computing Neuron*, R. M. Durbin, C. Miall, and G. J. Mitchison, eds., Chap. 4, pp. 54–72. Addison-Wesley, Wokingham.

Földiák, P. 1990. Forming sparse representations by local anti-Hebbian learning. *Biol. Cybernet.* **64**, 165–170.

Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernet.* **36**, 193–202.

Gross, C. G., and Mishkin, M. 1977. The neural basis of stimulus equivalence across retinal translation. In *Lateralization in the Nervous System*, S. Harnad, R. Doty, J. Jaynes, L. Goldstein, and G. Krauthamer, eds., pp. 109–122. Academic Press, New York.

Hinton, G. E. 1987. Learning translation invariant recognition in a massively parallel network. In *PARLE: Parallel Architectures and Languages Europe*, G. Goos and J. Hartmanis, eds., pp. 1–13. Lecture Notes in Computer Science, Springer-Verlag, Berlin.

Hubel, D. H., and Wiesel, T. N. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154.

Klopf, A. H. 1982. *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence.* Hemisphere, Washington, DC.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comp.* **1**, 541–551.

Meister, M., Wong, R. O. L., Baylor, D. A., and Shatz, C. J. 1990. Synchronous bursting activity in ganglion cells of the developing mammalian retina. *Invesli. Ophthalmol. Visual Sci.* (suppl.) **31**, 115.

Mitchison, G. J. 1991. Removing time variation with the anti-Hebbian synapse. *Neural Comp.*, in press.

Perrett, D. I., Rolls, E. T., and Caan, W. 1982. Visual neurons responsive to faces in the monkey temporal cortex. *Exp. Brain Res.* **47**, 329–342.

Rumelhart, D. E., and Zipser, D. 1985. Feature discovery by competitive learning. *Cog. Sci.* **9**, 75–112.

Spitzer, H., and Hochstein, S. 1985. A complex-cell receptive-field model. *J. Neurophysiol.* **53**, 1266–1286.

Sutton, R. S., and Barto, A. G. 1981. Toward a modern theory of adaptive networks: Expectation and prediction. *Psychol. Rev.* **88**, 135–170.

Webber, C. J. St. C. 1991. Self-organization of position- and deformation-tolerant neural representations. *Network* **2**, 43–61.

Received 20 September 1990; accepted 12 October 1990.