Proceedings of 1993 International Joint Conference on Neural Networks

Learning invariant responses to the natural transformations of objects[1]

Guy Wallis,[2] Edmund Rolls and Peter Földiák.

Oxford University, Department of Experimental Psychology,
South Parks Road, Oxford OX1 3UD, England.

Abstract

The primate visual system builds representations of objects which are invariant with respect to transforms such as translation, size, and eventually view, in a series of hierarchical cortical areas. To clarify how such a system might learn to recognise 'naturally' transformed objects, we are investigating a model of cortical visual processing which incorporates a number of features of the primate visual system. The model has a series of layers with convergence from a limited region of the preceding layer, and mutual inhibition over a short range within a layer. The feedforward connections between layers provide the inputs to competitive networks, each utilising a modified Hebb-like learning rule which encorporates a temporal trace of the preceding neuronal activity. The trace learning rule is aimed at enabling the neurons to learn transform invariant responses via experience of the real world, with its inherent spatio-temporal constraints. We show that the model can learn to produce translation-invariant responses.

1 Introduction

There is evidence that over a series of cortical processing stages, the visual system of primates produces a representation of objects which shows invariance with respect to, for example, translation, size, and view, as shown by recordings from single neurons in the temporal lobe (see Rolls 1992; Tanaka 1988). In a recent paper, Rolls (1992) reviews much of this work, with specific regard to those cells responsive to faces, and goes on to advance a theory for how these neurons could acquire their transform independent selectivity. It is this analysis which forms the basis for the network described here.

Fundamental elements of Rolls'(1992) hypothesis are:

- A series of competitive networks, organised in hierarchical layers, exhibiting mutual inhibition over a short range within each layer.
- A convergent series of connections from a localised population of cells in preceding layers to each cell of the following layer, thus allowing the receptive field size of cells to increase through the visual processing areas or layers.
- A modified Hebb-like learning rule incorporating a temporal trace of each cell's previous activity, which, it is suggested, will enable the neurons to learn transform invariances.

To clarify the reasoning behind the third point consider the situation in which a single neuron is strongly activated by a stimulus within a real world object. The trace of this neuron's activation will then gradually decay over a time period in the order of 0.5s, say. If, during this

[1] The research was supported by a grant from the Human Frontier Science Program, and by the Oxford IRC in Brain and Behaviour.

[2] G. Wallis received a SERC studentship during the course of this work.

limited time window, the net is presented with a transformed version of the original stimulus then not only will the initially active afferent synapses modify, but so also will the synapses activated by this transformed version of this stimulus. In this way the cell will learn to respond to either appearance of the original stimulus. The cell will not, however, tend to make spurious links across stimuli that are part of different objects because of the unlikelihood of one object consistently following another. A possible biological basis for this temporal trace could lie in the persistent firing of neurons for 300-500ms observed after presentations of stimuli for as little as 16 ms (Rolls et al, 1993), or alternatively, in the fact that NMDA receptor activated channels remain activated for periods of up to several hundred milliseconds (Rolls, 1992). What follows is a summary of the work carried out on a network architecture which has been simulated to investigate these hypotheses.

2 Theory of Learning

The idea that incorporating a trace of cell activity could aid the learning of natural transforms of objects was first discussed in detail by Földiák (Földiák 1991). The learning rule used here is similar to Földiák's, and can be summarised as follows:

$$\Delta w_{ij}{}^{(t)} = \alpha \overline{y}_i{}^{(t)} . x_j \quad : \quad \sum_j w_{ij}{}^2 = 1 \ for\ each\ i^{th}\ neuron$$

and

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

where x_j is the j^{th} input to the neuron, y_i is the output of the i^{th} neuron, w_{ij} is the j^{th} weight on the i^{th} neuron, η governs the relative influence of the trace and the new input (typically $0.4 - 0.6$), and $\overline{y}_i{}^{(t)}$ represents the value of the i^{th} cell's trace at time t. Note that in this simulation neuronal learning is bounded by normalisation of each cell's dendritic weight vector. An alternative, more biologically relevant implementation, using a local weight bounding operation, has in part been explored using a version of the Oja update rule (Oja 1982; Kohonen 1984).

3 The Network

3.1 Architecture

The forward connections to a cell in one layer are derived from a topologically corresponding region of the preceding layer, using a gaussian distribution of connection probabilities to determine the exact neurons in the preceding layer to which connections are made. This schema is constrained to preclude the repeated connection of any cells. Each cell receives 50 connections from the 32x32 cells of the preceeding layer, with a 67% probability that a connection comes from within 4 cells of the distribution centre. Fig.1 shows the general convergent network architecture used, and fig.2 Rolls' proposal for convergence in the primate visual system. Within each layer, lateral inhibition between neurons has a radius of effect just greater than the radius of feedforward convergence just defined. The lateral inhibition is simulated via a linear local contrast enhancing filter active on each neuron. (Note that this differs from the global 'winner-take-all' paradigm implemented by Földiák, 1991). The cell activation is then passed through a non-linear cell output activation function.

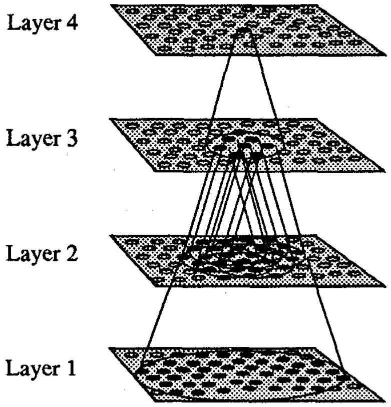

Figure 1: Hierachical network structure.

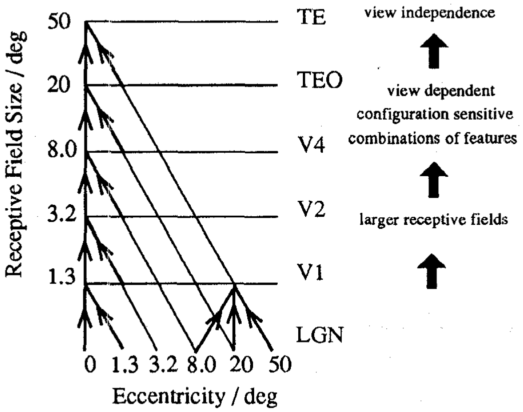

Figure 2: Convergence in the visual system.

3.2 Network Input

In order that the results of the simulation might be made more relevant to understanding processing in higher cortical visual areas, the inputs to layer 1 come from a separate input layer which provides an approximation to the encoding found in visual area 1 (V1) of the primate visual system. These response characteristics are provided by a series of spatially tuned filters with image contrast sensitivities chosen to accord with the general tuning profiles observed in the simple cells of V1. Currently, only even symmetric - bar detecting - filter shapes are used. The precise filter shapes were computed by weighting the difference of two gaussians by a third orthogonal gaussian according to the following formula:[3]

$$\Gamma_{xy}[f,\theta,\rho] \;=\; \rho\left(\frac{(\sqrt{3}+1)}{\sqrt{3}}\,e^{-\left(\frac{x\cos\theta+y\sin\theta}{\sqrt{2}/f}\right)^2} - \frac{1}{\sqrt{3}}e^{-\left(\frac{x\cos\theta+y\sin\theta}{\sqrt{6}/f}\right)^2}\right)e^{-\left(\frac{x\sin\theta+y\cos\theta}{3\sqrt{2}/f}\right)^2}$$

where f is the filter spatial frequency (in the range 0.0625 to 0.25 pixels^{-1} over four octaves), θ is the filter orientation (0^o to 135^o over four orientations), and ρ is the sign of the filter i.e. ± 1. Cells of layer 1 receive a topologically consistent, localised, random selection of the filter responses in the input layer, under the constraint that each cell samples every filter spatial frequency and receives a constant number of inputs.

4 Analysis of learning

In order to test the network a set of three non-orthogonal stimuli, based upon probable 3-D edge cues (such as a 'T' shape), was constructed. During training these stimuli were chosen in random sequence to be swept across the 'retina' of the network, a total of 1000 times. In order to assess the characteristics of the cells within the net, a two-way analysis of variance was performed on the set of responses of each cell, with one factor being the stimulus type and the other the position of the stimulus on the 'retina'. A high F ratio for stimulus type (F_s), and low F ratio for stimulus position (F_p) would imply that a cell had learned a position invariant representation of the stimuli. The discrimination factor of a particular cell was then simply gauged as the ratio $\frac{F_s}{F_p}$, (a measure for ranking at least the most invariant cells[4]).

[3] We thank Dr. R. Watt, of Stirling University, for assistance with the implementation of this filter scheme.

[4] We are grateful to Dr. F. Marriott for his statistical advice.

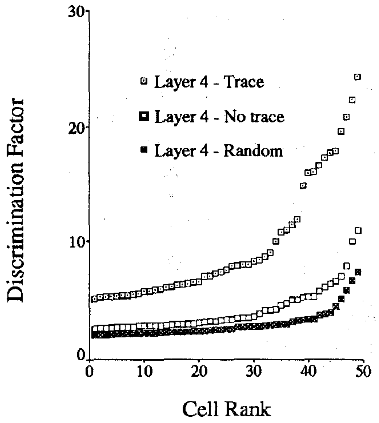

Figure 3: Comparison of network discrimination.

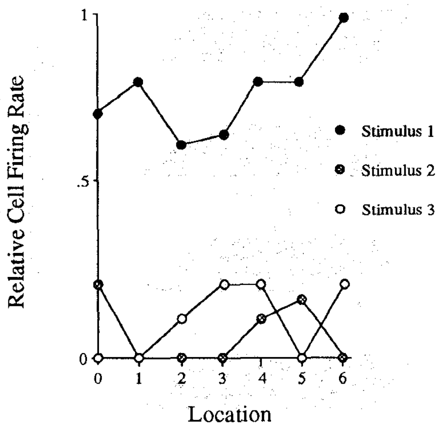

Figure 4: Cell showing stimulus selectivity.

To assess the utility of the trace learning rule, nets trained with the trace rule were compared with nets trained without the trace rule and with untrained nets (with the initial random weights). The result of the simulations, illustrated in fig.3, show that networks trained with the trace learning rule do have neurons with much higher values of $\frac{F_s}{F_p}$. An example of the responses of one such cell are illustrated in fig.4. Similar position invariant encoding has been demonstrated for a stimulus set consisting of faces.

5 Conclusions

The results described in this paper show that the proposed learning mechanism and neural architecture can produce cells with responses selective for stimulus type with considerable position invariance. Although only translation invariance with a limited number of stimuli has been investigated here, in future investigations it will be important to include the use of a much larger stimulus set. It would also be of interest to investigate invariance learning for other 'natural' object image transforms. The ability of the network to be trained with natural scenes may also help to advance our understanding of encoding in the visual system.

References

Földiák, P.(1991). Learning invariance from transformation sequences. *Neural Computation 3(2), 194-200.*

Kohonen, T.(1984). Self-organization and associative memory. *Pub. Springer-Verlag.*

Oja, E.(1982). A simplified neuron model as a principal component analyser. *Jnl. Math. Biol. 15, 267-273.*

Rolls, E.(1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical areas. *Phil. Trans. Roy. Society London Ser. B 335, 11-21.*

Rolls, E. and Tovee, M.(1993). Processing speed in the cerebral cortex, and the neurophysiology of visual masking. *In preparation.*

Tanaka, K. Saito, H. Fukada, Y. and Moriya, M.(1991) Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Jnl. of Neurophysiology 66(1), 170-189.*