during the maggot's explosive jump. The ants subsequently failed to relocate their prey which landed several centimetres away. Thus, the unpredictable nature and force of jumping may prove to be an effective escape mechanism for fruit-fly larvae.

Given their total lack of appendages and their soft hydraulic-based skeletal system, it is surprising that maggots jump[5]. Indeed, jumping maggots appear to be the only known examples of jumping by soft-bodied legless organisms. Other examples of legless jumping are found in hard-bodied legged invertebrates, but these employ alternative mechanisms to jump (click beetles[6], springtails[7], bristletails[8] and millipedes[9]). The jumping performances of maggots and click beetles, together with two other well known jumpers, are compared in Table 1.

In light of the ability of these maggots to jump, a caterpillar-like body plan which uses hydraulic-based skeletal and locomotory systems obviously does not preclude high-speed locomotion[1]. In addition, the reported restrictions on stride length[1] are overcome to a degree by caterpillars (Geometridae) and leeches looping instead of crawling. Clearly, the hydraulic body of caterpillar-like organisms is more versatile than was previously believed. □

Received 27 August; accepted 9 October 1991.

1. Casey, T. M. *Science* **252,** 112–114 (1991).
2. Wainwright, S. A. *Axis and Circumference: The Cylindrical Shape of Plants and Animals* (Harvard University Press, Cambridge, Massachusetts, 1988).
3. Wong, T. T. Y., Ramadan, M. M., McInnis, D. O. & Mochizuki, N. *J. econ. Entomol.* **83,** 779–783 (1990).
4. Wong, T. T. Y., Mochizuki, N. & Nishimoto, J. I. *Environ. Entomol.* **13,** 140–145 (1984).
5. Wille, *J. Zool. Jb. Allg. Zool.* **39,** 301–320 (1922).
6. Evans, M. E. G. *J. Zool.* **167,** 319–336 (1972).
7. Manton, S. M. *The Arthropods* (Clarendon, Oxford, 1977).
8. Evans, M. E. G. *J. Zool.* **176,** 49–65 (1975).
9. Evans, M. E. G. & Blower, J. G. *Nature* **246,** 427–428 (1973).
10. Schmidt-Nielsen, K. *Animal Physiology: Adaptation and Environment* 4th edn (Cambridge University Press, UK, 1990).
11. Alexander, R. McN. *Animal Mechanics* 2nd edn (Blackwell, Oxford, 1983).
12. Phillips, V. T. *Memoirs of the American Entomological Society No. 12* (Philadelphia, 1946).
13. Shadwick, R. E. & Gosline, J. M. *J. exp. Biol.* **114,** 259–284 (1985).
14. Vogel, S. *Life's Devices: The Physical World of Animals and Plants* (Princeton University Press, NJ, 1988).
15. Alexander, R. McN. *J. theor. Biol.* **124,** 97–110 (1987).

ACKNOWLEDGEMENTS. I thank P. Maitland for assistance, H. Benze for translations, D. Dawson for the loan of video equipment and M. Villet for identifying ants.

Self-organizing neural network that discovers surfaces in random-dot stereograms

Suzanna Becker & Geoffrey E. Hinton

Department of Computer Science, University of Toronto, 10 King's College Road, Toronto M5S 1A4, Canada

THE standard form of back-propagation learning[1] is implausible as a model of perceptual learning because it requires an external teacher to specify the desired output of the network. We show how the external teacher can be replaced by internally derived teaching signals. These signals are generated by using the assumption that different parts of the perceptual input have common causes in the external world. Small modules that look at separate but related parts of the perceptual input discover these common causes by striving to produce outputs that agree with each other (Fig. 1*a*). The modules may look at different modalities (such as vision and touch), or the same modality at different times (for example, the consecutive two-dimensional views of a rotating three-dimensional object), or even spatially adjacent parts of the same image. Our simulations show that when our learning procedure is applied to adjacent patches of two-dimensional images, it allows a neural network that has no prior knowledge of the third dimension to discover depth in random dot stereograms of curved surfaces.

The simplest way to get the outputs of two modules to agree is to use the squared difference between the outputs as a cost function, and to adjust the weights (connection strengths) in each module so as to minimize this cost. Unfortunately, this usually causes each module to produce the same constant output that is unaffected by the input to the module and therefore conveys no information about it. What we want is for the outputs

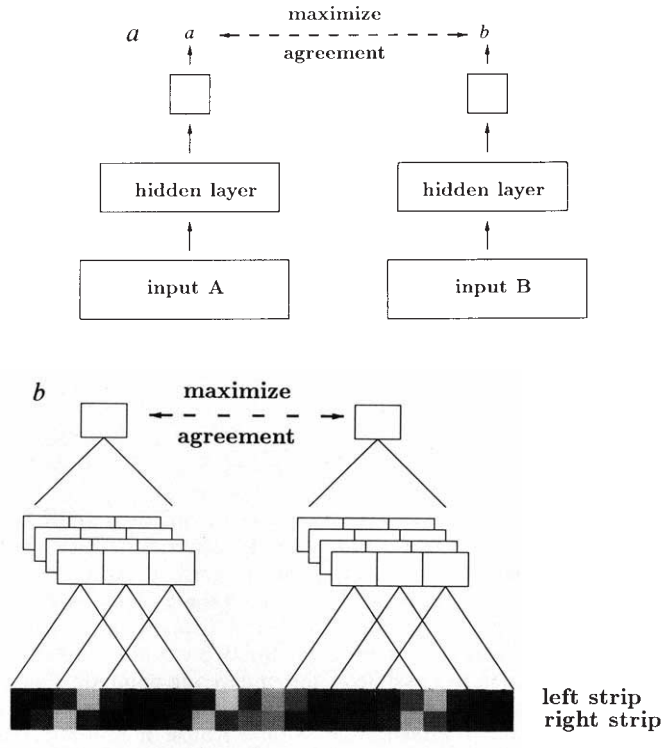

FIG. 1 *a*, Two modules that look at separate but related parts of the perceptual input and attempt to produce the same output. By differentiating some measure of the agreement between the outputs, we can generate a teaching signal that can be back-propagated through each module to compute how to change the connection strengths. In our initial simulations the output units are linear and the hidden units use the logistic nonlinearity $y = (1+\exp(-\mathbf{w}\cdot\mathbf{s}-b))^{-1}$ where $\mathbf{w}$ is the weight vector of a unit, $\mathbf{s}$ is the input vector and b is the bias. *b*, Two modules that receive input from corresponding parts of stereo images. The input pattern is an example of a random-dot steogram of a surface which is curved in depth. A strip of curved surface is generated by fitting a smooth curve (a cubic spline) to four or more control points whose depths are chosen at random. Random dots are scattered sparsely on the surface strip, and a pair of stereo images is made by taking two slightly different projections. The projections are filtered through a gaussian and sampled at evenly spaced sample points. In our images, disparity ranges continuously from −1 to +1 image pixels. The sample values in corresponding patches of the two images are used as the inputs to a module. For simplicity, we use image strips one pixel high. Each module has one layer of hidden units, organized into clusters of three feature detectors (units in the same cluster are shown in the same plane). Within a cluster, the feature detectors have the same weights but the receptive fields of neighbouring units (indicated by triangles) are translated by two pixels. Because the three hidden units within a plane are constrained to learn translated versions of the same feature, we have four planes to allow four different features to be learned. The learning algorithm adjusts the weights in each module to maximize the agreement measure, *I*. The derivative of *I* provides error signals that are propagated backwards[1] through the layers of each module. After each weight update, we average corresponding weights in all the modules to enforce the constraint that every module computes exactly the same function of its input vector. We also average corresponding weights of units in a cluster of three feature detectors. These equality constraints reduce the number of free parameters that must be learned which speeds the learning and limits the ability of the system to discover spurious correlations in the data that are caused by the limited size of the training set. To speed the learning further, we used a conjugate gradient optimization method for updating the weights in the modules.

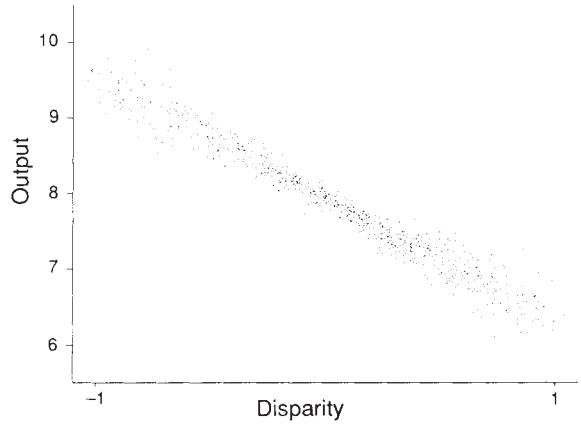

FIG. 2 The output of a module as a function of the depth for 1,000 test cases of random-dot stereograms of curves surfaces. Pairs of corresponding hidden units were trained to maximize agreement using 20 iterations of steepest descent learning. Then 10 conjugate gradient iterations (typically about 250 function evaluations) were used to maximize agreement between the outputs of neighbouring modules.

of two modules to agree closely (that is, to have a small expected squared difference) relative to how much they both vary as the input is varied. When this happens, the two modules must be responding to something that is common to their two inputs. In the special case when the outputs, a and b, of the two modules are scalars, a good measure of agreement is

$$I = 0.5 \log \frac{V(a+b)}{V(a-b)} \quad (1)$$

where V is the variance over the training cases. If a and b are both versions of the same underlying gaussian signal that have been corrupted by independent gaussian noise, it can be shown that I is the mutual information between the underlying signal and the average of a and b (S.B. and G.E.H., unpublished results). By maximizing I we force the two modules to extract as pure a version as possible of the underlying common signal.

Just as back-propagation can be viewed as a multilayer nonlinear extension of linear regression because it minimizes a squared error measure, the procedure we present here can be seen as a multilayer nonlinear extension of statistical methods such as canonical correlation or alternating conditional expectation[2] because it maximizes the normalized covariation between the outputs.

In the following simulations, we used 1,000 training cases of random-dot stereo images such as those shown in Fig. 1*b*. The real-valued depth (relative to the plane of fixation) of each patch of the surface gives rise to a disparity between intensity peaks in the left and right images. The disparity is the only property that is coherent across each stereo image. We trained a network that contained 10 modules using the architecture shown in Fig. 1*b*. Each module had one real-valued output and learned to extract depth by trying to maximize agreement with the outputs of immediately adjacent modules.

With random initial weights, the derivatives of I are tiny so learning is very slow. We therefore introduced an initial learning phase in which we trained corresponding pairs of units in the hidden layers of neighbouring modules to agree. It is impossible for these units to achieve very high agreement because an intermediate layer is required to extract depth properly. Also there is no pressure for different hidden units in the same module to discover different features. But even modest depth tuning of the hidden units makes the subsequent learning much easier for the output units. During the second phase of the learning we trained the output units of neighbouring modules to agree. Derivatives of this agreement were propagated backwards through the hidden layers to fine-tune the hidden units to be as useful as possible to the output units. Output units became accurately tuned to depth (Fig. 2).

So far, we have used a very simple model of coherence in which an underlying parameter at one location is assumed to be roughly equal to the parameter at a neighbouring location. This model is fine for the depth of fronto-parallel surfaces but it is far from the best model of slanted or curved surfaces.

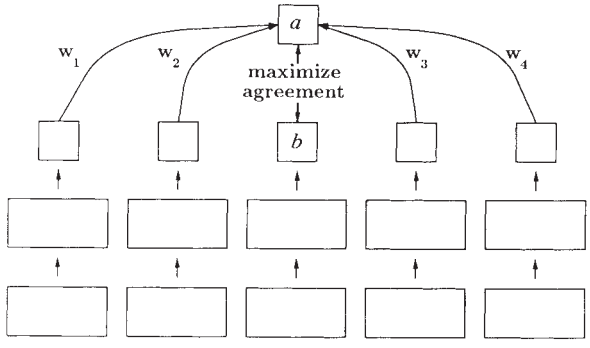

FIG. 3 The architecture of the network used for discovering locally detectable parameters that are linear combinations of nearby parameters. The network consists of multiple copies of modules like those in Fig. 2 plus a layer of interpolating units that are used for predicting the locally extracted parameter from several nearby parameters. We actually used 10 modules and the central six modules tried to maximize agreement between their outputs and contextually predicted values. We used weight averaging to constrain the interpolating function to be identical for all modules. We tested this model on several image ensembles with varying amounts of curvature, by varying the number of control points used to generate the cubic spline surfaces. After having been trained for 50 conjugate gradient iterations, the four weights learned for the interpolating function were: Two control points (CP): 0.256, 0.266, 0.258, 0.265; three CP: 0.018, 0.516, 0.531, 0.011; four CP: −0.147, 0.675, 0.656, −0.131; five CP: −0.244, 0.734, 0.746, −0.256. As the curvature increases, a characteristic pattern emerges in which positive weights are given to inputs coming from the immediately adjacent modules, and smaller negative weights are given to inputs coming from the more distant neighbours. The output of the interpolating units is similar to the response profile shown in Fig. 3, but even more finely depth-tuned. When we increase the difficulty of the learning problem by making the random-dot densities greater, the network learns a qualitatively similar interpolating function to the ones shown above but with smaller weights. This is a sensible solution because when predicting a depth from noisy estimates of nearby depths, the noise amplification is proportional to the sum of the squares of the weights.

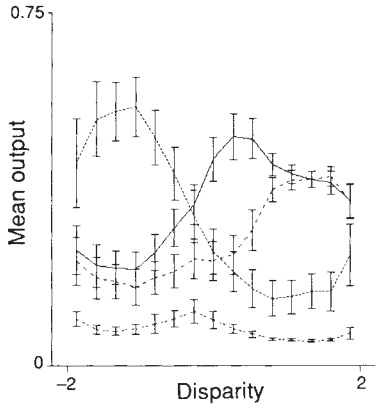

FIG. 4 The depth-tuning curves of the four competing output units of a module. The curves are averages over 1,000 test cases of random-dot stereograms of curved surfaces. The training and test patterns were created in the same manner as described in Fig. 2, but this time the disparity ranged continuously from −2 to +2 pixels. The agreement measure used for the learning is a modified version of the function defined in equation (1). (S.B. and G.E.H., manuscript in preparation).

Fortunately, we can use a far more general model of coherence in which the parameter at one location is assumed to be an unknown linear function of the parameters at nearby locations. The particular linear function that is appropriate can be learned by the network.

We used a network of the type shown in Fig. 3. The depth computed locally by a module, *b*, was compared with the depth predicted by a linear combination of the outputs of nearby modules, *a*, and the network tried to maximize the agreement between *a* and *b*. The contextual prediction, *a*, was produced by computing a weighted sum of the outputs of two adjacent modules on either side. The interpolating weights used in this sum, and all the other weights in the network, were adjusted so as to maximize agreement between locally computed and contextually predicted depths. To speed the learning, we first trained the lower layers of the network as before, so that agreement was maximized between neighbouring locally computed outputs. This made it easier to learn good interpolating weights. When the network was trained on stereograms of cubic surfaces, it learned interpolating weights of −0.147, 0.675, 0.656, −0.131. Given noise-free estimates of local depth, the optimal linear interpolator for a cubic surface is −0.167, 0.667, 0.667, −0.167.

There are many possible variations of this learning procedure. We assumed that the depth of a patch is represented by the activity level of a single linear unit. An alternative representation, which seems to be used by binocular cortical cells[3], is to have a population of units each of which is tuned to a range of depths. This representation has a number of advantages, including the ability to represent uncertainty by uniformity of activity across the population. Our measure of agreement can be modified to produce such population codes (Fig. 4).

We have described the learning procedure for modules which each have a single real-valued output. For modules with several real-valued outputs, the natural way to generalize the objective function is to replace the variance by the determinant of the covariance matrix. When this version of the procedure is applied to an ensemble of images of the same two-dimensional shape with different poses (that is, different positions, sizes and orientations), it extracts the four parameters of the pose, because different local fragments of the object all agree on the pose of the whole object[4].

The procedure we have described was designed to eliminate the need for an external teacher, but it may also overcome another weakness of supervised learning procedures. A global external teaching signal causes interdependencies between all the weights in the network which leads to slow learning in large networks. If pairs of small modules can generate their own teaching signals by trying to maximize agreement, many pairs can learn in parallel without interference. Also, the outputs of modules that have already learned can become the inputs to other modules that look for more complex or longer-range coherence. This should make learning much faster in very large networks. □

Received 10 May; accepted 4 October 1991.

1. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. *Nature* **323,** 533–536 (1986).
2. Hastie, T. J. & Tibshirani, R. J. *Generalized Additive Models* (Chapman and Hall, London, 1990).
3. Lehky S. R. & Sejnowski, T. J. *J. Neurosci.* **10,** 2281–2299 (1990).
4. Zemel, R. S. & Hinton, G. E. in *Advances in Neural Information Processing Systems* Vol. 3 (eds Lippman, R. P., Moody, J. E. & Touretzky, D. S.) 299–305 (Morgan Kaufmann, San Mateo, CA, 1991).

ACKNOWLEDGEMENTS. We thank P. Brown, F. Crick, A. Jepson, B. Pearlmutter and M. Revow. This research was supported by the Canadian Institute for Advanced Research and the Ontario Information Technology Research Center.

Rapid-time-course miniature and evoked excitatory currents at cerebellar synapses *in situ*

R. Angus Silver, Stephen F. Traynelis & Stuart G. Cull-Candy

Department of Pharmacology, University College London, Gower Street, London WC1E 6BT, UK

NEUROTRANSMISSION from mossy fibre terminals onto cerebellar granule cells is almost certainly mediated by L-glutamate[1,2]. By taking advantage of the small soma size, limited number of processes and short dendrite length of granule cells, we have obtained high-resolution recordings of spontaneous miniature excitatory postsynaptic currents (m.e.p.s.cs) and evoked currents in thin cerebellar slices[3]. Miniature currents have a similar time-course and pharmacology to evoked currents and consist of an exceptionally fast non-NMDA (*N*-methyl-D-aspartate) component (measured rise-time, 200 μs; estimated pre-filtered rise-time < 100 μs; decay time constant, $\tau = 1.0$ ms), followed by 50 pS NMDA channel openings that are directly resolvable. We could find no evidence for the recent proposal that miniature currents in granule cells are mediated solely by NMDA channels with a novel time course[4]. The non-NMDA receptor component of m.e.p.s.cs has a skewed amplitude distribution, which suggests potential complications for quantal analysis. The difference in time course between the m.e.p.s.cs reported here and other synaptic currents in the brain[5–8] could reflect differences in synaptic function or electrotonic filtering; the relative contribution of these possibilities has yet to be established.

Mossy fibre–granule cell synaptic transmission invariably produces dual component excitatory postsynaptic currents (e.p.s.cs) (Fig. 1*a*, *c*). The fast component (mean ± s.e.m., 1,700 ± 300 pS; $n = 22$) is reversibly blocked by the non-NMDA receptor antagonists CNQX (6-cyano-7-nitro-quinoxaline-2,3-dione) and NBQX (6-nitro-7-sulphamoyl-benzo(f)quinoxaline-2,3-dione)[9] (3–10 μM; $n = 9$; Fig. 1*b*, *d*), whereas the slow component can be reversibly and selectively inhibited by the NMDA receptor antagonists APV[9] (10–20 μM; $n = 14$, Fig. 1*a*, *b*) and 7-chlorokynurenate (5–10 μM; $n = 6$), and by Mg^{2+} (1 mM, at negative potentials). The pharmacologically isolated fast non-NMDA component (Fig. 1*a*, *c*, inset) had a 10–90% rise-time of 410 ± 30 μs and decay time constant of 1.3 ± 0.1 ms (monotonically rising events, $n = 15$ cells); this contrasted with the characteristically slow rise (10–90% rise-time 9 ± 1 ms; $n = 7$) and decay