NEURAL NETWORKS AS A MODEL FOR VISUAL PERCEPTION: WHAT IS LACKING?

Rolf P. Würtz

Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany

Abstract

A central mystery of visual perception is the classical problem of invariant object recognition: Different appearances of an object can be perceived as "the same", despite, e.g., changes in position or illumination, distortions, or partial occlusion by other objects.

This article reports on a recent email discussion over the question whether a neural network can learn the simplest of these invariances, i.e. generalize over the position of a pattern on the input layer, including the author's view on what "learning shift-invariance" could mean. That definition leaves the problem unsolved. A similar problem is the one of learning to detect symmetries present in an input pattern. It has been solved by a standard neural network requiring some 70000 input examples. Both leave some doubt if backpropagation learning is a realistic model for perceptual processes.

Abandoning the view that a stimulus-response system showing the desired behavior must be learned from scratch, yields a radically different solution. Perception can be seen as an active process that rapidly converges from some initial state to an ordered state, which in itself codes for a percept. As an example, I will present a solution to the visual correspondence problem, which greatly alleviates both problems mentioned above.

1. Introduction

Artificial neural networks have introduced a radically new way of computer programming by providing simple and convincing ways for a system to learn concepts from examples. The importance of this can not be overestimated, and

---

Revised version of a paper presented at the 14th Annual Workshop of the European Society for the Study of Cognitive Systems, Saint Maximin la Sainte-Baume, 17–20 April, 1996.

---

there are many practical applications which can even be sold for good money (one of the major quality criteria of our time).

From a naive viewpoint, it is enticing to apply this tool to the difficult task of constructing computational models of perception. After all, the brain consists of nothing but a lot of neurons with an even larger number of connections. Given sufficient computing power all the mysterious properties of perception should be exhibited by a network which – after the presentation of many examples – develops from a simple unstructured state to a system with cognitive capabilities.

That view can be challenged from various perspectives. First of all, most neural network models have the structure of simple stimulus-response schemes, where the same input always produces the same output unless the internal state has been changed by learning. This ignores important psychological parameters like readiness to show or repress a response and other internal factors that obey their own laws. These considerations lead beyond the scope of this article, because they point to the tough issues of consciousness, namely, how are the results of cerebral computation coded, and how do conscious results (i.e., conscious information about the environment) differ from unconscious ones (e.g. intermediate results or raw sensory data), which are, even with effort, not accessible. These are interesting and important questions, however, I will refrain from any speculations about the issue. The reason is that I see no way of getting a grip on it, not qualitatively and even less quantitatively enough for computer modeling.

2. **Shift-invariance**

Let us now return to more technical and more tractable questions. A central mystery of visual perception is the classical problem of invariant object recognition: Different appearances of an object can be perceived as "the same", despite, e.g., changes in position or illumination, distortions, or partial occlusion by other objects. I wish to report on some aspects of an email discussion of the question "Can neural networks learn shift-invariance?", which took place in 1996. This is the simplest of invariances, namely the recognition of a stimulus independently of its position on the retina. It was triggered by a request for a proof for the "generally held view" that the answer was negative.

2.1. *Can Neural Networks Learn Shift-invariance?*

The problem is the following: There are N possible patterns at P possible positions. The complete set of P·N possibilities must be divided into a training set and a test set. (Patterns with symmetries are ignored for the sake of simplicity). The network has succeeded when it has learned the correct classification of the N patterns independent of their position. There is no disagreement that neural networks are capable of performing shift-invariant recognition. A good example is the neocognitron (Fukushima et al., 1983). However, the necessary structure is already built into the neocognitron before it starts learning.

It can be shown that for N=1 and a training set of size P-1 there are examples where no local learning rule can generalize to the shift-invariant solution. That means that the concept of shift-invariance can not generalize to positions where the pattern has never been seen. This is probably more than one should expect, anyway.

The challenge for a neural net to learn shift-invariance can be defined as follows. Beginning from tabula rasa, the network is presented *one* pattern in *all* possible positions to learn the concept of shift-invariance. For practical reasons, more than one pattern may be required, but I would insist that shift invariance has to be learned from a small number ($S \ll P$) of the possible patterns.

After having learned shift-invariance that way the network should be able to learn new patterns at a *single* position and then recognize them in an invariant way in *any* position. Again, I would allow a small number of positions. It is granted that the network is *now* a structured one, but the structure has been learned from examples.

A good step into this direction is a network by Hinton (1987), which actually achieves the desired generalization. His parameters are N=16, P=12. Every pattern is trained at 10 (random) positions. So the number of training examples is $0.83 \cdot P \cdot N$, the number of test examples to which the network generalizes is $0.17 \cdot P \cdot N$.

This gets a little awkward for larger values of N and P. The task as outlined above would allow only $S \cdot (P+N-1)$ training examples. Something like S=3 should be appropriate, S=1 desirable. Then the network should generalize and recognize all $P \cdot N$ examples correctly. Note that there is no objection to the choice of parameters in (Hinton, 1987) but to the scaling behavior for larger parameter values. The network must have seen all patterns in almost all possible positions to do the generalization.

To the author's knowledge the goal of learning shift-invariance from a training set of size $S \times (P+N)$ with a small S has not been reached yet. So the question whether a standard feed-forward neural network with a local learning rule can learn the concept of shift invariance remains open. The concept itself seems so important that a failure to do so will challenge the claim that the respective classes of models can achieve a reasonable description of perception.

The problem can, of course, be solved by extending the allowed structures appropriately. Giles and Maxwell (1987) showed that higher order networks containing, e.g. $\Sigma\Pi$-units are able to learn shift invariance. These are neurons that have multiplicative as well as the usual additive synapses, the former being useful for *gating* connections according to a global transformation. to date, it seems difficult to find experimental evidence for such gating synapses, sometimes also called three-axon terminals. In Section , a quite different extension of neural networks will be discussed.

2.2. *Does the Human Visual System Exhibit Shift-invariance?*

As an aside, the problem could also be solved the other way around. Thus, an interesting thread of the discussion asked the question whether the human visual system shows shift-invariance at all. The psychophysical data are not perfectly conclusive at the moment. Biedermann and Cooper (1991) find complete invariance in priming experiments. In these experiments, people were shown line drawings of familiar objects and, in a second run, had to identify the objects shown in the first one. The fact that they do better in the second run is called object priming, and the experiments proved that reaction times as well as error rates were independent of the position of the second presentation. Presentation times were to short to adjust fixation.

Experiments done by Nazir and O'Regan (1990) used completely unknown and meaningless patterns. They had to be learned at one position and recognized at another one. The results show that error rates and reaction times were significantly worse if the position of presentation was changed. This seems to be evidence for imperfect shift-invariance in the case of unknown objects, although Irving Biedermann commented that Nazir and O'Regan do find strong shift-invariance under more thoroughly controlled conditions. Anyway, there seems to be agreement about the fact that recognizing patterns at shifted positions can cause distinctly more *effort* than at the ones, where the pattern was learned.

3. **Dynamic-link Networks**

A major simplification made by standard neural network theory is that neurons do their processing instantaneously, without any internal dynamics, and that the time course of the activity is not important. This causes theoretical difficulties – it has been suggested that, e.g., the well-known binding problem can only be overcome if the information content of the temporal structure on a fine time scale is exploited by the brain (von der Malsburg, 1981). While there may be feature detectors for cows and goats and brown and purple, it is most unlikely that there are specialized detectors for purple cows, before the first instance is encountered. It is also implausible that they develop instantly on this very first encounter. So somehow the detectors must be rapidly bound together (and just as rapidly be cut apart) during perception. This is not trivial, because pure detector activities can not distinguish between a scene with a purple cow and a green goat and another one with the colors reversed.

A radical solutions to the binding problem was proposed by von der Malsburg (1981,1985,1995) . He postulated that pairs of cells which have a physical connection and, therefore, a synaptic weight, which can be changed by learning over long time intervals, can also have a short-term weight, i.e. a property that can vary on a time scale of a couple of milliseconds. These *dynamic links* also develop by dynamics (which are formally similar to Hebbian learning) and also influence the activities of the cells involved. This mechanism can bind “purple” to “cow” and “green” to “goat” and thus resolve the above mentioned ambiguity. It is important to notice that the concept of dynamic links extends the two time scales present in conventional neural networks (immediate vs. learning) by a third one that should be synchronous with the time scale of perception. Natural systems usually operate on a continuum of time scales, the slowest one being biological evolution.

3.1. *Computability of Brain States*

Another aspect might be worth considering. A standard feed-forward neural network represents a simple function from an input pattern to an output pattern. With known (computable) weights, this function is obviously computable in the sense that a discrete machine can calculate its output. With the introduction of recurrence the timing of each neuron begins to matter. With dynamic links the whole system can only be described by a complex system of differential equations.

Such differential equations perform, in principle, computations with infinite precision. As a consequence, it might be impossible to simulate them on a Turing machine in the sense of computing their development up to any desired precision. An example of a simple linear partial differential equation for which this non-computability can actually be proved, has been presented by Pour-El and Richards (1981) . Bournez and Cosnard (1996) study the problem in much more detail, and their result is that many differential equations are computationally more powerful than Turing machines. If one accepts the view that the brain can be modeled by a system of differential equations these results open a possibility that essential properties of the brain can, in principle, not be modeled using a digital computer, no matter how parallel its architecture might be. Put differently, it *might* be the case that the brain can be modeled by neural networks in a usual way, but those networks can not be simulated.

3.2. *Shift- and Deformation-invariance in Dynamic-link Networks*

The architecture outlined in the beginning of this Section can accommodate shift-invariance in the following way. First, the notion of *correspondence* is introduced. Given two images I_1 and I_2 of the same object, it must be decided which point in I_1 corresponds to which in I_2 in the sense that both are projections from the *same* point on the physical object. This is called the *correspondence problem*. If it is solved for sufficiently many points, shift- and deformation-invariance is easy to achieve. It suffices to compare local features at all pairs of corresponding points, and if the sum (or average) of these local similarities is high the objects are similar. Such a comparison can be done for several memorized objects and the object with the highest similarity value is then recognized.

The solution of the correspondence problem, however, is by no means trivial. If small distortions and rotations are allowed, the corresponding points in two identical sheets of unstructured white paper are *nowhere* uniquely determined. Up to rotation by multiples of 90° they are determined only at the four corners, all others can vary in a considerable range. In the presence of structure, the problem remains that local features are not unique, i.e. non-corresponding points frequently carry the same features. Introducing more complex and thus less ambiguous features can help, but these are usually more sensitive to distortions. The only way to resolve these ambiguities is taking the spatial arrangement of points into account in addition to the feature similarities.

In the following, I will briefly describe a class of dynamic-link networks that can solve the correspondence problem for human faces (Würtz, 1995; Wiskott and von der Malsburg, 1996; Wiskott, 1996; Würtz and von der Malsburg, in preparation). Image and model (the actual and the memorized face) are both represented by a sheet of neurons each of which carries a local feature vector (a hypercolumn of simple cells) with it. Both sheets are fully interconnected by dynamic links, i.e. each neuron has a connection to all neurons in the other layer. Internally, the layers are connected by fixed short-range exciting and long-range inhibiting connections. These connections allow only a restricted set of activity patterns. For appropriate parameter values the only possible patterns are *one* island of activity in a sea of inactive neurons. The position of this so-called activity *blob* is determined by irregularities of the input, e.g. by means of the dynamic links between the layers, its size is controlled by the parameters of the intralayer connections.

Some simple further machinery causes the blob in each layer to move about its layer. The crucial step is now the development of the dynamic links. The strength of each link grows if both of the cells it connects are activated by the blobs *and* if their feature vectors are similar. Growth of one link can only occur at the expense of other links connected to the same cell. In the beginning, the blobs move independently. With growing links, the probability increases that the blobs are at corresponding positions, because there the feature similarity will be high. Once this has happened the process is self-amplifying: The correct pairs are always active at the same time *and* have high feature similarity, so their links will grow, non-corresponding pairs are active at different times, so their links will shrink due to the growth of the other ones. This system converges to a one-one-mapping between the layers that reflects the correct correspondences.

Variations of this system with different emphasis are described by von der Malsburg (1988), Lades et al. (1993), Konen and Vorbrüggen (1993), Würtz (1995), Würtz and von der Malsburg (1996, in preparation), Wiskott (1996) and Wiskott and von der Malsburg (1996). Konen and Vorbrüggen (1993) and Wiskott and von der Malsburg (1996) extend the above system to a complete object recognition system by means of competition between several object layers. Würtz (1995) and Würtz and von der Malsburg (1996, in preparation) reduce the sequential processing time by a coarse-to-fine version of this principle.

Another variation of this system can possibly offer an interesting solution to the binding problem if one assumes that memory traces can *induce* the synchronicity required for the emergence of an ordered state. This idea has been developed in considerable detail by Phillips and Singer (1998), a possible application to visual feature extraction is proposed by Würtz (1998) . However, a working system that would demonstrate its capabilities is still lacking.

3.3. *Symmetry recognition in Dynamic-link Networks*

Another successful example for the application of dynamic-link Networks is the recognition of symmetries in an input pattern. This problem has been solved by Sejnowski et al. (1986) using a Boltzmann machine. Their network was able to learn to classify, e.g., 10×10 patterns for vertical, horizontal, or diagonal symmetry after the presentation of 70000 examples. Even then, the success rate was about 70%.

The main reason for the necessity of such an exorbitant training phase is that their network has no idea that feature correspondences are important — it is too general for the problem at hand. It could, with the same effort, learn symmetries between a pattern and an *arbitrary permutation* between the pixels. Clearly, this is nothing a living brain would easily do. Thus, solving the correspondence problem by a dynamic-link network is the way to go.

Konen and von der Malsburg (1993) interconnected two layers containing copies of the input pattern by dynamic links which develop by a similar mechanism as the one described in Section . They usually find a one-one mapping between the two copies of the pattern in as few as 40 activation cycles. A simple perceptron of 3 output and 18 hidden units reads the temporal correlations between the pattern layers and learns the correct classification from the presentation of one or two training examples per symmetry class.

Dynamic-link Networks are designed to produce smooth mappings between an image and a model. This general smoothness constraint is shared by many of the transformations that are important in vision (i.e., translations, perspective changes, size and orientation changes and internal deformations), although the networks can cope with cases where the smoothness is violated locally (due to, e.g., occlusion). This makes them ideally suited for model matching and object recognition.

A substantial disadvantage from the computational point of view is that their processing is inherently partly sequential, so they require a certain amount of time for operation. Note that this is the time to build up a percept, not the time it is held in memory. On the other hand, recognition in the brain also requires time and this time is even dependent on the input data. In that sense, classical neural networks fail to model reaction time, one of the most important observables in psychophysics.

3.4. *Biological Relevance of Dynamic-link Networks*

After having seen some of the power of dynamic-link networks the question naturally arises whether this is at all realistic in the biological sense. Strangely enough, the actual physical basis of synaptic strength remains mysterious, the experimental data about activity-dependent strengthening of synapses (long-term potentiation or LTP) and their weakening (long-term depression or LTD) is obtained only indirectly from the spike activities of the cells. Currently, synaptic strength can not be measured directly.

For many years experimental evidence for changes of synaptic strength at the time scale of several milliseconds was not available and the whole concept of dynamic links remained speculative. This situation has changed due to the findings by Makram and Tsodyks (1996) and Abbott et al. (1997). They present evidence that the activity of the postsynaptic cell can very rapidly influence the synapses at their dendrite. This gives enormous computational power to a single neuron, due to the complexity of the dendritic tree. More data must be accumulated before the function of these capacities will become clear, but rapid synaptic plasticity which is guided by coherent activity in pre- and postsynaptic neuron is now a definite possibility.

4. **Conclusion**

I have challenged the usefulness of standard neural networks for modeling perception from a technical and a fundamental point of view. To illustrate the technical problems, I have chosen the classical question of how invariant recognition can be learned. It has been shown that those problems can be overcome by extending the structures allowed in these systems, particularly by shifting attention to a new time scale. The proposal of dynamic-link networks begins to be well-covered by neurobiological findings. The combination of conceptual power and biological plausibility makes them very promising candidates for the construction of realistic models of perception.

Acknowledgements

In Section , I have used arguments from an email discussion on the “connectionist” mailing list. Short of knowing a proper way to cite such an event I would like to thank Irving Biedermann, Georg Dorffner, Jerry Feldman, Lee Giles, Geoffrey Hinton, Eric Postma, and Laurenz Wiskott for their unintentional contributions to this paper. Further thanks go to G.J. Dalenoort and Jan C. Vorbrüggen for the improvements they suggested.

References

Abbott, L., Varela, J., Sen, K., and Nelson, S. (1997), Synaptic depression and cortical gain control, *Science*, 275:220–224.

Biedermann, I. and Cooper, E. E. (1991), Evidence for complete translational and reflectional invariance in visual object priming, *Perception*, 20:585–593.

Bournez, O. and Cosnard, M. (1996), On the computational power and super-Turing capabilities of dynamical systems, Technical Report NC-TR-96-005, Royal Holloway, University of London.

Fukushima, K., Miyake, S., and Ito, T. (1983), Neocognitron: a neural network model for a mechanism of visual pattern recognition, *IEEE Trans. SMC*, 13:826–834.

Giles, C. and Maxwell, T. (1987), Learning, invariance, and generalization in high-order neural networks, *Applied Optics*, 26(23):4972–4981.

Hinton, G. (1987), Learning translation invariant recognition in massively parallel networks, In Goos, G. and Hartmanis, J., editors, *PARLE Parallel Architectures and Languages Europe*, number 258 in Lecture Notes in Computer Science, pages 1–13, Springer.

Konen, W. and von der Malsburg, C. (1993), Learning to generalize from single examples in the dynamic link architecture, *Neural Computation*, 5:719–735.

Konen, W. and Vorbrüggen, J. (1993), Applying dynamic link matching to object recognition in real world images, In Gielen, S., editor, *Proceedings of the International Conference on Artificial Neural Networks*, North-Holland, Amsterdam.

Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., and Konen, W. (1993), Distortion invariant object recognition in the dynamic link architecture, *IEEE Transactions on Computers*, 42(3):300–311.

Macphail, E. M. (1987), The comparative psychology of intelligence, *Behavioral and Brain Sciences*, pages 645–695.

Makram, H. and Tsodyks, M. (1996), Redistribution of synaptic efficacy between neocortical pyramidal neurons, *Nature*, 382:807–810.

von der Malsburg, C. (1981), The correlation theory of brain function, Technical report, Max-Planck-Institute for Biophysical Chemistry, Postfach 2841, Göttingen, FRG, Reprinted 1994 in: Schulten, K., van Hemmen, H.J. (eds.), *Models of Neural Networks*, Vol. 2, Springer.

von der Malsburg, C. (1985), Nervous structures with dynamical links, *Ber. Bunsenges. Phys. Chem.*, 89:703–710.

von der Malsburg, C. (1988), Pattern recognition by labeled graph matching, *Neural Networks*, 1:141–148.

von der Malsburg, C. (1995), Binding in models of perception and brain function, *Current Opinion in Neurobiology*, 5:520–526.

Nazir, T. A. and O'Regan, J. K. (1990), Some results on translation invariance in the human visual system, *Spatial Vision*, 5(2):81–100.

Phillips, W. A. and Singer, W. (1997), In search of common foundations for cortical processing, *Behavioral and Brain Sciences*, in press.

Pour-El, M. B. and Richards, I. (1981), The wave equation with computable initial data such that its unique solution is not computable, *Advances in Mathematics*, 39:215–239.

Sejnowski, T., Kienker, P., and Hinton, G. (1986), Learning symmetry groups with hidden units: Beyond the perceptron, *Physica*, 22D:260–275,

Wiskott, L. (1996), *Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis*, Reihe Physik, Verlag Harri Deutsch, Thun, Frankfurt am Main.

Wiskott, L. and von der Malsburg, C. (1996), Face recognition by dynamic link matching, In Sirosh, J., Miikkulainen, R., and Choe, Y., editors, *Lateral Interactions in the Cortex: Structure and Function*, The UTCS Neural Networks Research Group, Austin, TX, Electronic book, ISBN 0-9647060-0-8, http://www.cs.utexas.edu/users/nn/web-pubs/htmlbook96.

Würtz, R. P. (1995), *Multilayer Dynamic Link Networks for Establishing Image Point Correspondences and Visual Object Recognition*, volume 41 of *Reihe Physik*, Verlag Harri Deutsch, Thun, Frankfurt am Main.

Würtz, R. P. (1998), Context dependent feature groups, a proposal for object representation, *Behavioral and Brain Sciences*, in press.

Würtz, R. P. and von der Malsburg, C. (1996), A hierarchical dynamic link network to solve the visual correspondence problem, *Perception*, 25(Suppl.):27–28.

Würtz, R. P. and von der Malsburg, C. (in preparation), A hierarchical dynamic link network for correspondence maps between image points.

Manuscript received: 5-3-1998

Adress author:
Institut für Neuroinformatik
Ruhr-Universität Bochum
D–44780 Bochum
Germany
E-mail:Rolf.Wuertz@neuroinformatik.ruhr-uni-bochum.de