Perspectives on neural network models and their relevance to neurobiology

1989 J. Phys. A: Math. Gen. 22 1959

(http://iopscience.iop.org/0305-4470/22/12/001)

J. Phys. A: Math. Gen. 22 (1989) 1959-1968. Printed in the UK

Perspectives on neural network models and their relevance to neurobiology†

Gérard Toulouse

Département de Physique, Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France

Received 25 January 1989

Foreword

During the past two years, whenever asked to introduce neural networks for an audience of physicists, I used to start with a sketch of history, admittedly simplified and organised from a physicist's vantage point—a broader and fairer context was given afterwards—in order to reconstruct the main steps, leading from the genesis of the topic to its present state. The name of Elizabeth Gardner was associated with the most recent such step, the seventh one.

Here is the list of individuals and their contributions.

(1) W S McCulloch and W Pitts (1943) for the description of a neuron as a binary, all-or-none, element, and for showing that networks of such simple elements can perform logical computations.

(2) D O Hebb (1949) for the notion that a percept or a concept is represented in the brain by a cell assembly and the suggestion that learning occurs by modification of synaptic efficacies.

(3) B G Cragg and H V Temperley (1954) for the analogy between neural network persistent activity and the collective states of coupled magnetic dipoles.

(4) W A Little (1974), for the analogy between noise and temperature, that paved half the way towards thermodynamics.

(5) J J Hopfield (1982, 1984), for the study of models of content-addressable memory with a concept of 'energy', that completed the linkage to thermodynamics, and for the analogy to spin glasses.

(6) D J Amit, H Gutfreund and H Sompolinsky (1985) for showing that a class of such models was amenable to exact treatment, and yielded striking results.

(7) E Gardner (1987) for the systematic exploration of the space of couplings, a novel approach in statistical mechanics, and for the consequent opening of new vistas in this science.

As early as the days in late 1986 and early 1987 at Ecole Normale, when Elizabeth presented to us her preliminary results (later published as 'The space of interactions

† This paper was originally given as an informal summary talk at the end of the Bat-Sheva Seminar held in Jerusalem from 24 May to 3 June 1988. No formal references are given but all scientists mentioned, except some in an historical context, were contributors to the Seminar. For publication details, see the bibliography at the end of this paper.

in neural network models'), I felt convinced that this was a major conceptual breakthrough, as is now obvious from all the results that have blossomed along her approach (e.g. the recent Virasoro work on categorisation). She was aware of this recognition and asked me to write in support of her application for a lectureship at Edinburgh: my report was sent on April 12. Little did I know, then, that the remarkable achievements of this modest, gentle young woman were offered to us in the urgency of a fight against death. In retrospect, the brilliant creativity and the admirable dignity of her last years are the best apology for pure research and the ethics of knowledge.

The year-round programme at the Institute for Advanced Studies of Jerusalem, in which Elizabeth participated during two months in the winter of 1988, culminated in a two-week workshop on 'Neural Network Models and their Relevance to Neurobiology', for which I was asked to draw a summary and perspectives. This final session took place a couple of weeks before Elizabeth's passing, which news came as a shock for all of us.

In agreement with the colleagues who asked me to contribute a written version of this talk (Hanoch Gutfreund kindly suggested some improvements), I have decided to keep the casual spirit of the oral presentation. What follows is a snapshot, dated 3 June 1988.

1. Introduction

Drawing perspectives at the end of a conference is a difficult exercise. Indeed, predictions of scientific activity, for the short term, are bound to be close to linear extrapolation, and thus trivial; for the long-term, they are likely to fall into banalities, and thus be boring. For the middle-term, predictions may be influential, and thus pernicious; too easily misleading, they are potentially dangerous.

Frankly, I have been subjected to many final conference talks in my career (this is the first time I inflict one) and as a whole they did not leave me with very strong impressions—with a few exceptions, one of which will be mentioned later on.

A story and an outline

A man had two wives, one old and one young. The old wife would pick the black hair from his beard, so that he would look older; the young wife would pick off the white hair, so that he would look younger. In the end, the man was left with no beard.

Take this as an allegory for confrontations between biologists and physicists. Let the man be 'brain science'; his beard, the 'models'; the old wife, biology; the young wife, physics. The biologists are eager to take away the unbiological elements in order to make the models more realistic; the physicists are prone to discarding the unessential parameters, to make the models more soluble. After our two weeks of vigorous debate, a superficial observer might slide toward the conclusion that no neural modelling is left at the physics-biology interface. But let us be a little more perceptive with, first, a survey of some basic issues:

the number of cell types, in neural systems;
the properties of a single neuron;
the synapses, as sites for learning;

the important scales, and the appropriate measurements; and, second, a discussion centred on two debates:

what is the neural code? and, more specifically, the confrontation: computing-by-dynamic-flow or computing-with-attractors?

the choice between 'realistic' *ab initio* models versus 'simplified', idealised, models.

2. Survey of basic issues

2.1. How many cell types?

We have heard several estimates for a mammalian cortex ranging from 2 (V Braitenberg and I White) to 10^7 (T Bullock). This does not imply a disagreement about facts, since the criteria for class identification may vary in degree of coarseness. Our question will be more precisely phrased: how many cell types should be included in a neural network model, in order to achieve some biological relevance?

For clarification, it should be emphasised that we are interested here in anatomical differences, as distinct from merely functional ones. In some sense, each neuron is different from any other, because their geometrical location and their connections differ. But this kind of difference will be found in any random structure.

In the basic models of statistical mechanics, for heterogeneous (spin glass) as well as for homogeneous systems, there is traditionally only one type of constituent element (e.g. the binary Ising spin). For simplicity, this feature has been carried through into many early neural network models. However, let me remind you of D Lehmann's talk, where he reported a significant and promising step toward the biological findings for the cortex, namely the inclusion of two neuron types (one excitatory such as the pyramidal cell, the other inhibitory such as the stellate cell).

On the frontier of model investigations, other kinds of differences between cell types clearly deserve attention, such as differences between receptor neurons (coupled to the inputs) and processing neurons (hidden from the outside), and differences in the characteristics of time constants (inputs integration, refractory periods, response delays).

2.2. How many parameters to describe a single neuron?

The neurobiologists, as they refine their tools (anatomy, physiology, chemistry), strive to explore the complexity of a single neuron (I Segev). As a result of their successful endeavours, the list of parameters characterising a single neuron becomes an ever increasing one. This makes even more crucial, for the theoretician—be he or she of biological or physical origin—to separate the details (to be discarded) from the essentials (to be included). Of course, the separation is not intangible and will in general depend on the level of comprehension that is looked for.

Much of the excitement, during recent years came from the surprise that models with very simple formal neurons were producing a wealth of results on the storage properties of distributed memories. But remember how primitive present theories are, when no consensus is achieved on some fundamental questions.

(*a*) *Are the spikes relevant*? Spikes may be totally irrelevant to functional behaviour of the cortex, argued T Bullock, as he advocated the study of compound electric fields, averaging over large neural populations. The firing frequency of a neuron is, generally, the relevant variable, claimed J Hopfield (in an analogy with electrical currents, a single spike might be as irrelevant as the trajectory of a single electron), in support of the graded deterministic, model (as defined in his 1984 paper). But the spike noise deserves attention because it is often quite helpful functionally, stressed D Amit, H Gutfreund and H Sompolinsky, advancing along the road of the discrete stochastic model (as defined by Hopfield in 1982). And, in complete contrast with the above, biologists or physicists, M Abeles presented his synfire chain theory, where the spike structure is absolutely essential, due to synchronicity effects ignored in other theoretical formulations.

(*b*) *Low activities of a neuron.* In the most naive interpretation of the theory of computation-with-attractors (from Hebb's cell assembly formulation to the Hopfield-like high-feedback models), a neuron involved in a persistent state (an attractor) is supposed to fire either strongly (close to maximum firing) or weakly (close to minimal firing).

Now, although occasional bursts are recorded in the cortex, the analysis shows that they do not fit the criteria to be relevant as evidence for persistent activity states (they are much too rare and fleeting). The bursts being ruled out, at least in the cortex, what can be observed (E Vaadia) is the switching of a neuron between two low-activity states (say, one with 5 spikes per second, the other with 20). However, from the theoretical side, it has proved difficult to devise neural network models that present this feature (two different low-activity states), with sufficient structural stability.

During this conference, an illuminating discussion between J Hopfield and H Sompolinsky has suggested that the input-output response curve of the neuron (firing frequency f plotted against synaptic inputs, or membrane potential U) may be crucial for this effect. Namely if, everything else remains unchanged, the neuron response curve is changed from the standard smooth sigmoid form (figure 1(*a*)) to a form with a sharp threshold, or more generally one with a region of large slope (figure 1(*b*)); the property of switching between two stable states of low activity ensues naturally.

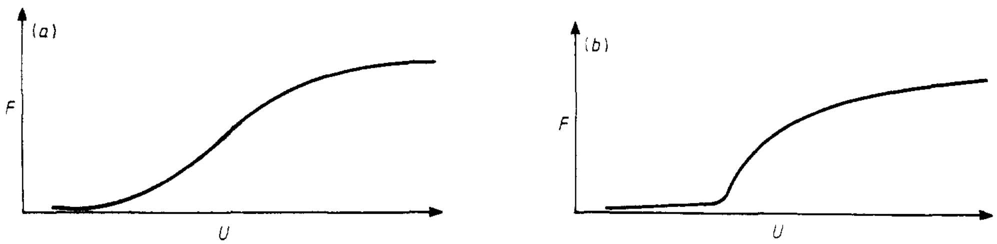

Figure 1. Neuron response curves plotted as firing frequency against membrane potential, showing (*a*) a smooth sigmoid curve and (*b*) a curve with a region of large slope.

Although further analysis is needed in this example, it is illustrative of the way by which neural network studies can help to extract the crucial parameters from the inessential details, and stimulate new experimental activity (here, in order to check the neuron response curves) and new theoretical activity (here, in order to explore the consequences of this modification within the set of neuron characteristics).

(*c*) *Time constants.* Is delayed response an attribute of a significant fraction of neurons? This is an example of an increase in formal neuron complexity. Recent model studies have shown the virtues of cells with delayed response for the production and recognition of temporal sequences. As a result, the biologists are now encouraged to look harder for such cells.

Is the neuron a coherence detector? Questions of timing between synaptic inputs along the dendritic geometry, and synchrony effects in general, are ignored or poorly addressed in the standard models.

(*d*) *Thresholds as dynamical variables.* In addition to the usual dynamical variables, i.e. neural activities and synaptic efficacies (P Peretto), recent work (D Horn) explores the possible role of thresholds and suggests their inclusion as relevant variables.

2.3. Synapses

Along with Hebb ideas, the synapses are generally considered as the main sites for modifications due to learning.

(*a*) *Does the change occur on the presynaptic side or on the postsynaptic side*? Research in Aplysia, and long-term potentiation in the hippocampus, as reviewed by J Byrne and H Segal, have fostered the 'pre' hypothesis, whereas the 'post' hypothesis has been much elaborated by considerations on the role of dendritic spines in the cortex (V Braitenberg).

At issue is the existence of a Hebbian-like mechanism for synaptic modification, because the convergence of signals from the presynaptic and postsynaptic neurons will take place more naturally on the postsynaptic membrane (figure 2).

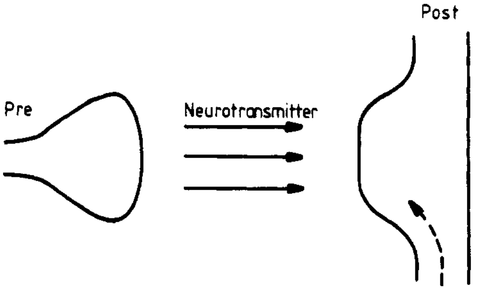

Figure 2. Convergence of signals from presynaptic and postsynaptic neurons.

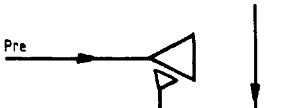

Figure 3. The Aplysia-like circuitry for presynaptic change.

If, however, a presynaptic mechanism takes place, then the information from the postsynaptic neuron must arrive on the presynaptic membrane via a retrograde signal through the junction (but so far there is no evidence for it), or via *ad hoc* circuitry (but this seems too contrived to be general) (figure 3).

(*b*) *Short-term memory versus long-term memory.* If two different learning mechanisms are implied, as suggested by invertebrate studies, and protein synthesis is involved in long-term memory, then how will the molecule be targeted from the unique protein production centre (the neuron nucleus) to one specific synapse among the many? That remains a mystery.

(*c*) *Accuracy in the synaptic efficacies.* How dependent are the storage properties claimed by various models on a precise monitoring of the synaptic efficacies? And what kind of accuracy control is biologically plausible?

2.4. Relevant scales and measurements

What we would like to measure is not necessarily what we have access to, and vice versa. At present, on two limited windows are working thousands of scientists: the 'population activity people' and the 'spike recorders', who do not communicate much with each other.

(*a*) *Compound fields of large populations.* T Bullock has presented the case for electroencephalographic recordings. His was a minority view, in this meeting, and it is not currently leading to active confrontation with neural network theories, because the models so far generally lack any geometry, and are therefore unable to bring testable predictions to guide the experiments.

(*b*) *Single neuron.* The grandmother-cell hypothesis (H Barlow) is next to impossible to check, whether true or not, because the likelihood of recording from 'the' specialised neuron, involved in a particular recognition task, is almost zero. At present, hardly testable, this thesis is unpopular.

(*c*) *Cell assembly.* This is the conceptual framework—'circuits compute, not populations'—favoured by neural network modellists and also by neurobiologists involved in single-unit or multi-unit recordings (Gerstein). It is useful to distinguish two simple extremes in the ensemble of models: the feedforward 'perceptron' and the feedback 'ganglion'.

The 'perceptrons' (E Domany) are layered structures, where information flows from the first layer (input) to the last layer (output). To this scheme belongs also the synfire chain theory (M Abeles), with an additional suggestion on how feedforward processing might emerge in *a priori* non-layered structures.

The 'ganglions' are defined here as richly interconnected structures where the inputs govern the initial network state. Processing of the information occurs via the internal dynamics of the network, which eventually settles into an attractor. This persistent state then contains the output of the computation.

Let us examine in more detail the debate between the two approaches.

3. Debate: computing-by-dynamic-flow versus computing-with-attractors

3.1. A few reminders

(*a*) On a scale going from early to late processing (S Hochstein and D Sagi), one can distinguish several stages: sensory, preprocessing (this stage may be shorter in olfaction than in vision or audition), recognition, attention.

(*b*) Brain anatomy suggests that there is always an unambiguous sense of flow, that can be determined from the layer-to-layer projections, which are qualitatively different between forward and backward direction (D Van Essen). This is true in visual cortex, inferotemporal cortex, hippocampus, cerebellum, ...

From anatomy also, note this quantitative fact stressed by V Braitenberg: in 1 mm^2 of cortex (which is a biologically relevant scale), there are as many synapses originating from outside cells as from inside cells. Is this figure accidental, or is it the result of some trend toward maximal complexity (in the sense of defining a system equally distant from the two simpler architectures of the pure feedforward perceptron and the pure fully connected ganglion)?

(*c*) So far, the experimental evidence for persistent states comes *either* low down, from very primitive circuits (such as invertebrate central pattern generators), *or* high up, from event-related recordings (such as delayed template matching) in tasks involving short-term memory and attention (where the observability may be due to a glueing of assemblies into superassemblies, increasing the likelihood that a randomly selected neuron takes part in the task).

3.2. The issue of typicality

It seems to me that most physicists would be happy enough if their models of computing-with-attractors had some sort of limited validity—say in the higher attentive tasks. But a biologist like Abeles insists on discovering 'the' neural code, namely the code used generally, in most activities of the cortex. Remember also the neat protest of S Hochstein: 'where is the division between feature detectors (preprocessing) and neural nets (processing)? It seems always further than where we record!'

3.3. Some candidates for the neural code

It is commonly agreed that at the sensory periphery, the firing frequency of a neuron codes for stimulus intensity.

By contrast, in association areas, sharply different views are held. Many people think that the firing frequency of a neuron codes for the probability that some proposition is true. An earlier-mentioned consequence of this view is that individual spikes become irrelevant, like single electrons in electric wires.

The brace in figure 4(*a*) emphasises a time interval when the neuron firing has become more frequent, and therefore significant. Note that the neuron (or the neurophysiologist recording from it) may know when it (or he) is doing something meaningful, from the observable change in the firing activity.

However, in the synfire chain scheme (M Abeles), isolated spikes are meaningful, and there is no way for the neuron to know which one of its spikes are significant (in figure 4(*b*) two 'relevant' spikes are marked, but no feature in the single-neuron

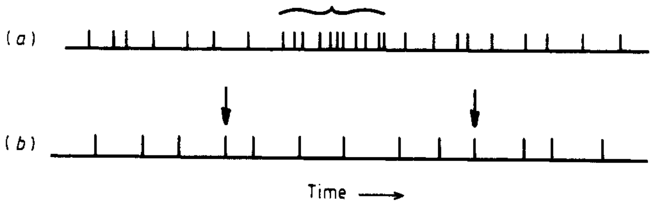

Figure 4. Neuron firing frequency coding (*a*) for stimulus intensity and (*b*) in the synfire chain scheme.

recording distinguishes those from the others). This obviously has implications for learning. How can meaningful learning take place at the proper synapses, if the information is not available locally (and if we are reluctant to use the escape hypothesis of higher supervisory neurons)?

3.4. Pattern recognition

Let us assume that pattern recognition is the 'typical' cortex activity. In contrast with the task-related states of E Vaadia, two experimental difficulties for microelectrode recordings appear: fewer neurons and smaller times.

Presumably, in a typical pattern recognition computation, much smaller cell assemblies are involved than in attention states. The time intervals may be between 30 ms and 200 ms (Abeles), instead of 10 s.

Note that it may be useful to distinguish between the computation time (time needed to reach the solution of the task) and the retention time (time during which the result of the computation is held available for further processing). In speech, it may be that the information on word recognition is passed on immediately to higher areas, and need not be retained as such, whereas information on the spatial environment, and everything that defines the mental 'frame' inside which the subject organises his or her activity, may require long retention times (it is perhaps the main virtue of the computing-with-attractors schemes to provide a simple solution for variable retention times, which is not so easy to explain in computing-by-dynamic-flow).

4. Digression on physics, biology and the neural code

There is a famous pronouncement by Sir Ernest Rutherford (1871-1937), who was a physicists' physicist, though he got a Nobel prize in chemistry. He said something like 'Science is physics, and (the rest is) stamp collecting'. At some times, during this seminar, I felt that Moshe Abeles was about to claim 'Neural science is biology, and (the rest is) model collecting'.

Indeed, till Rutherford's days, biology was diversity. Then, largely under the influence of physicists such as Delbrück, Schrödinger, Szilard, Gamow, Perutz, Luria and Crick, it turned into a search for universal laws. Remember the universal genetic code (despite exceptions recently found in paramecium, etc); the central dogma that genetic information flows from DNA to RNA to proteins (despite exceptions like retroviruses, etc); the selective mechanism of the immune system, 'one B-cell, one antibody'; and so on.

In these studies, many biologists have adopted the physics strategy of looking for the simplest system, on which universal features can be unravelled and crucial tests performed. Remember the bold saying 'Why study the elephant, when there is the bacteria?' Meanwhile, during the same period of time, much of the thrust of physics has turned from the study of fundamental laws into the exploration of emergent collective properties of matter. And this has led physics into a diversity of models and metaphors.

At this point, let me recall a personal souvenir. That was in 1973, near Göteborg, at a Nobel Symposium on Collective Properties of Physical Systems. In an after-dinner speech, the physicist Harry Suhl, inspired by the recent discovery of two superfluid phases in helium 3, both of which had been predicted by two conflicting theories, concluded 'Provided your model conserves energy and momentum, Nature will be kind enough to present a realisation for it'.

This philosophy may explain the flourishing of models and—even more remotely from realistic matter studies—of metaphors and scenarios, that are accumulated into repertoires in which you take your pick *a posteriori*, according to the observed behaviour of your system. Call it *ab termino* theory, in contradistinction with *ab initio*. This course of theorising has become common practice in hydrodynamics (e.g. approach to turbulence) and it is invading other fields.

Nonetheless, a number of physicists become dissatisfied with this state of affairs and are now turning towards biology, in order to recapture the harder confrontation with reality, as in the good old days. In some sense, these physicists turn to new biology out of faithfulness for old physics. So, I admit that Moshe has a point, and I would predict that, together with more continuation of statistical physics, and along with more exploration of toy models and of new computing devices (S Solla), there will be in the future a harder look and a stronger focus to study typical biological computations, such as pattern recognition in the cerebral cortex.

En attendant new scientific evidence, let me epitomise the main arguments in favour of computing-with-attractors.

On a set of neurons, there are several activity states, which are mutually exclusive: this forces a decision, that is taken by the network after weighing all the data.

Variable duration of attractor activity, i.e. variable retention time for a percept or a concept, is simply accounted for.

It is robust to neural degradation.

And now, the case for the synfire chain.

It is an efficient way for precise transport of information in an intricately connected network.

It appears to be a spontaneously emergent property of assemblies of 'coincidence detectors'.

It has no problem with low activity rates.

5. Debate: realistic *ab-initio* models versus idealised abstracted models

Despite some appearance to the contrary during the past days, when the colourfully displayed Caltech models of D Van Essen and J Bower aroused controversy with paper-and-pencil theoretical physicists, I believe that this is not a debate between biology and physics, but an internal debate within physics and within biology. Remember, in physics, all these *ab initio* models for band calculations in solids, molecular dynamics in liquids, models of atmosphere, etc. And in biology, the idealised Monod-Wyman-Changeux model for allosteric molecules that remains, 25 years later, an unavoidable reference.

Rather, the debate focuses on the proper use of computers, with their increasing computing power and graphic capacities, for heuristics. Use whatever approach you prefer: in the end, if your model is to be successful and adopted by others, even if you started playing with dozens of parameters, Occam's razor will prevail and what will remain is only the essentials (with, perhaps, the definition of several levels of analysis, using different sets of relevant parameters).

Let me try to draw a lesson from photosynthesis (credit for the idea should go to my colleague Pierre Joliot, but responsibility for mistakes in transcription is mine).

Photosynthesis appears to the observer as a horribly complicated superposition of three mechanisms, exquisitely optimised in every detail. Why is it so? Is this the result of an evolution process, stuck in a local minimum? Has this superposition a safety value? Whatever the answers, biological photosynthesis should not, emphatically not, be imitated globally for industrial photosynthesis, though inspiration can be drawn from the constituent mechanisms, after careful analysis and discrimination.

Now, is the photosynthesis lesson pertinent for the cerebral cortex? There are arguments for no and yes. No, because the cortex homogeneity suggests a general-purpose computer. Yes, because the multiplicity of areas (anatomy), the numerous visual illusions (psychophysics), suggest a manifold superposition of algorithms (a 'bag of tricks' rather than, say, a grand solution to such a grand problem as invariant recognition).

Sure, as stressed by T Sejnowsky, inspiration from biology does not imply putting in the models as many biological parameters as currently measured. Actually, the new developments in neural tissue cultures (D Kleinfeld) offer an attractive experimental procedure to analyse the relative influence of various neural parameters on their network properties.

As for directions in the future, let me be a cautious prophet and predict that the progress of brain science will be largely determined by the choices made by young people receiving multidisciplinary education in Caltech, Bell, Jerusalem, etc. Indeed, this year here has been a great opportunity for thoughts at the crossroads. Brera, yesh (Hebrew for: Choice, there is).

Bibliography

Two books have been written during the programme of the Jerusalem Institute, and the interested reader will find in them an adequate introduction to the subject and complementary surveys.

Amit D J 1989 *Modelling Brain Function* (Cambridge: Cambridge University Press)
Peretto P 1989 *The Modeling of Neural Networks* (Les Ulis: Editions de Physique)