PREPRINT N.899
9 Settembre 1992

Dipartimento di Fisica
Università di Roma "La Sapienza"
I.N.F.N. - Sezione di Roma

Belfast 25.6.1992
Second Irish Conference on Neural Networks

Roma-899-1992

Fermilab Library
0 1160 0018043 4

Cognitive-neuro-physiology: a empirical basis for neural modeling and cognitive psychology

Daniel J. Amit[1]
INFN, Sezione di Roma, Istituto di Fisica
Università di Roma, La Sapienza, P.le Aldo Moro, Roma

Network substrates for cognitive psychology

Networks of neuron-like elements have taken, in the last 10 years with the advent of PDP-1 and PDP-2, the center of the stage as a tool for capturing and/or producing dynamical phenomena of the complexity of cognitive systems. As such they have become a powerful source of models and phenomena for the cognitive psychologist and the linguist, to either explain intelligent behavior, or to produce new dynamical facts to be searched for in intelligent behavior. The strength of the new paradigm has been such as to withstand rather poignant objections by authorities such as Chomsky or Fodor.

Some of the attraction of the network (sometimes referred to as connectionist) program is in its innovative outlook. But its wide, and relatively uncritical, acceptance must be traced to the 'internal labels', e.g. neurons, synapses, learning etc. All of which have the flavor of brain, which one would expect to be a natural context for the study of brain phenomena. It is the impression, rather peculiar in hindsight, that very simple models, can produce and account for phenomena that-kindle our great curiosity just becuse we consider them so intrinsically and fundamentally complex.

The dominant impact on psychology has been that of feed-forward networks. The reason has probably been that along this route one can discuss relatively complex cognitive processes using technically simple tools. The psychologist, or the converted linguist, did not have to engage in evaluating physiological plausibility; analysis of dynamical systems; learning theory; compelxity of input-output relations.

The attractor outlook

An alternative approach, still a neural approach, has been that of attractror neural networks (ANN). It emphasized the collective features of large, relatively homogeneous, assemblies of interacting neurons. These collective properties, such as

[1]On leave of absence from Racah Institute of Physics

the maintenance of a selective spike activity pattern in absence of a stimulus, concentrate more on the emergent nature of the dynamics of the neural assembly than on the preconceived nature of the computation. As a consequence, this approach demonstrated the existence of the very unlikely phenomenon of the reverberation of selective spike distributions in a variety of patterns, which could, moreover, be shown to be controllable by the synaptic structure. This tied together two ends of the composite Hebbian paradigm: assembly reverberation and synaptic learning. Yet, the paucity of the computational achievment has left this development in the hands of theoretical physicists, in some dialogue with neuro-physiologists.

The purpose of this note, which is based on the technical results (Griniasti, Tsodyks, Amit, Neural Computation, 92, Amit and Tsodyks, in preparation) is to report a very exciting development in this dialogue, which brings computation into the realm of attractor networks, but in a rather unexpected fashion. The computation is found to be much less connected to our preconceived ideas about surface logic or arithmetic, and rather more to the actual process of learning. Much the divorce intuited by Von Neumann (quoted below). Its special attractions are in its very close and detailed contact with exeriment and in its immediate implications for cognitive psychology.

Some of the concepts and concerns involved in the program of modeling brain function by attractor dynamics have been sketched in the text presented for the first Royaumont meeting (May 1991), and in Amit (Modeling Brain Function. Cambridge, 1989). A summary of that summary is: Modeling brain function by attractor neural networks, emphasizes the dominant effect of feed-back in the dynamics of neural assemblies. Its other face, expresses the results of 'computation' by a selective enhanced activity of a subset of the neurons in the assembly, which persist in the absence of the stimulus provoking the computation. The possible results of a computation in an assembly are stored **passively** in the set of synaptic efficacies governing the dynamics. Additional attractors (potential outcomes of computations) can be included in the repertoire by a process of learning – synaptic modification. The selective activity provoked by a stimulus is an activation of one out of the collection of memorized (learned) passive attractors.

It is important to emphasize that this picture does not imply that the entire cortex is participating in this relaxation (attractor) dynamics. It was not implied in the original speculations of Hebb about assembly dynamics. Careful experiments (see e.g. Miyashita, 1988 Nature 331, p. 68) indicate that this type of cortical dynamics takes place on a very localized level, i.e. in columns of about $1mm^2$ of cortex. Even this caveat has a proviso. What seems to be localized is the selective structured activity in cortex. It may very well be, and I strongly believe it is, that the entire cortex is participating in an unstructured, very stable attractor, which maintains the background spontaneous activity, so essential for the functioning of the structured operation (e.g. Amit and Tsodyks, NETWORK 1992·Vol 3).

The picture implied in the attractor scenario is that the basic mechanism in

structured cortical function is the local relaxation into one of a variety of attractors (reverberations), selected by a stimulus. A given stimulus may provoke different retrieval processes in several local regions. The neural activity patterns, population codes, in individual modules communicate their choice to other modules, or to muscles and glands, by a feed-forward communication of information. How this set of cortical processes represents a computation in the surface language of mathematics, logic or psychology, may be, at most, vaguely perceived on the horizon. That is why it is so important to have intermediate levels of study, as independent as possible of preconceived notions of computation. Before discussing a proposal for such an intermediate level, let me recall a warning of John Von Neumann, which I find aging increasingly well:

> ...Thus the outward forms of *our* mathematics are not absolutely relevant from the point of view of evaluating what the mathematical or logical language *truly* used by the central nervous system is...the above remarks about reliability and logical and mathematical depth prove that whatever the system is, it cannot fail to differ considerably from what we consciously and explicitly consider as mathematics. [The Computer and the Brain, Yale 1954 p. 82, emphases in the original]

Cognitive-neuro-physiology

In fact, here I would like to use one experiment and one model to illustrate both a potential intermediate level of empirically accessible 'computation', as well as the fact that the interpretation of this experiment shows, surprisingly, that it is performed mostly in the process of learning. Then, I will try to indicate some potential imlications for cognitive science.

But still concerning preliminaries, I owe some explanation of the term 'cognitive-neuro-physiology'. It goes back to Miyashita's monkeys, which I discussed in my contribution to the first Royaumont meeting. This set of experiments are a culmination of a program started by Fuster and Niki in the early 70's. It consists in training monkeys to perform a certain task, which we may associate in some vague sense with rudimentary mental computation. Typical examples are: learning to recognize a set of visual, or somato-sensory, patterns. Then the task may be the comparison of pairs of these pattern stimuli, following a delay between the presentation of the two stimuli to be compared. This is the 'cognitive' part. The experimenters then insert electrodes in parts of the cortex which are supposed to be involved in the task. Neural activity, specific to the stimuli, is registered during the delay period. Abeles has lately started a set of experiments where the tasks are much more complex and deserving of the name computation.

Localising the region of stimulus specific responses and studying the relation of the activity distributions to the stimulating patterns one may attempt a phenomenology of cognitive-neuro-physiology. After all, the activity distributions provoked by a stimulus and persisting in its absence are in some sense a representation,

on the neural assembly level, of the outcome a local processing of each of the stimuli. What this processing may imply is, of course, a hard problem. It is complicated at both ends: in describing the transformation operated on the incoming stimulus in reaching the attractor representation, as well as in establishing the transformation required to reach a presumed output from the persistent activity distribution.

The power of the Miyashita monkey-program (inasmuch as monkeys are admitted into the study of cognition) is that it promises a sample cut of the gordian knot. It allows a study of the relationships between these internal representations and the learning process. Where by learning I mean the establishment of the synaptic structure which leads to the neural dynamics translating the stimuli into the attractors.

Conversion of temporal to spatial correlations

In the second Miyashita experiment (Nature 1988, Nature 335, p. 817) two monkeys are trained to recognize a set of visual patterns, which are presented in a fixed order (temporal correlations). Following the training, the monkeys are presented, in a random sequence, pairs of patterns separated by delays as long as 16 seconds. They are trained to react differentially depending on whether the second pattern was or saw not the same as the first. Single electrode spike recordings, during the delay period, of a large sample of neurons in a localized column of AVT cortex, turn out to be fully reproducible. These recording have all the signatures of attractors in a many-body interacting system (see e.g. Amit 1992 NETWORK, for discussion). These specific delay activity distributions are restricted, for all stimuli described, to a column of $1mm^2$, within the AVT region.

The neural spike distributions during the delay period, in which no visual stimulus is present, is a representation of the processing performed on the relevant stimulus in that particular part of cortex. We have not confronted, so far, this empirical concept of representation with the recent wave of anti-representation criticism in the community of cognitive philosophy. The internal representations for different memorized stimuli, expressed in terms of spike activity distributions, are then found to be correlated, despite the fact that care has been taken to exclude correlations among the visual stimuli. In fact, the two monkeys are trained with the same sets of patterns which are presented in different orders. Moreover, one may reasonably suppose that the types of computer generated visual patterns used by Miyashita have no selective or existential meaning for the monkey, which might have led to correlated representations (I owe this observation to G. Toulouse). The findings of Miyashita are that the correlation between the internal representations of two different stimuli depend on the separation of the two visual patterns in the (fixed) sequence used during training (see Fig. 1). Quantitatively, it is found that the attractor (reverberation) representations are significantly correlated up to a separation of 5 (**five**) in the training sequence.

Note that this discovery concerns the relation between two internal representations, irrespective of how each stimulus is represented, or how the representation

relates to an eventual reaction. This is the cognitive-neuro-physiological cut proposed.

Before referring to the theoretical interpretation, there is a very important conclusion to be drawn from this experiment: the computation which involves the conversion of the temporal correlations in the training sequence into a spatial correlation of the corresponding attractors, is carried out in the learning phase. Its outcome is stored in the synaptic arrangement which stabilizes the attractors. The excitement of one of these correlated attractors is merely an activation of a known computational result, stored passively in the synapses. After all the attractor is activated by a single visual stimulus, and the patterns with which it should be correlated are not present.

A simple model

The surprising result of these experiments is accompanied by yet another surprise (Griniasti, Tsodyks, Amit Neural Computation 92):

- What the synapses have to memorize, in order for the attractors to be correlated to a distance of 5 in the training sequence, is a relation between contiguous stimuli in the sequence **only**.

The correlations of different attractors vanish **exactly** beyond a separation of 5 (see Fig 1b), when the dynamics of the network is represented by ± 1 variables, *a la* Hopfield. This result is robust to changes in the parameter connecting the consecutive patterns in the synapses. The result has been checked for more realistic internal dynamics of the neurons, including the dynamics of integrate-and-fire neurons. The result is not as spectacularly neat as in the simple ± 1 case. But, even in this noisy and quasi-realistic situation, the attractors produced show a dependence of their correlations on the separation of the stimuli in the training sequence. The correlation function has an essentially invariant form in entire intervals of the single parameter specifying the memorized connection of consecutive patterns, the parameter a, below. Then, crossing a certain value, the form changes abruptly, and a new interval of invariant correlation appears. The number of attractors strongly correlated increases from interval to interval with increasing value of the parameter.

A technicality

[To be specific, a technical interlude: the simple case is a system of neurons that can take at any moment the value $S_i(t) = \pm 1$, storing uncorrelated patterns $\xi_i^\mu = \pm 1$, in a matrix of synaptic connections The

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \left[\xi_i^\mu \xi_j^\mu + a(\xi_i^{\mu+1} \xi_j^\mu + \xi_i^\mu \xi_j^{\mu+1}) \right], \quad (1)$$

where i and j are neuron numbers and p is the total number of memorized stimuli presented in trining, labeled by μ, which are considered to compose an ordered sequence. The ordering within the sequence corresponds to the fixed ordering of visual images in the

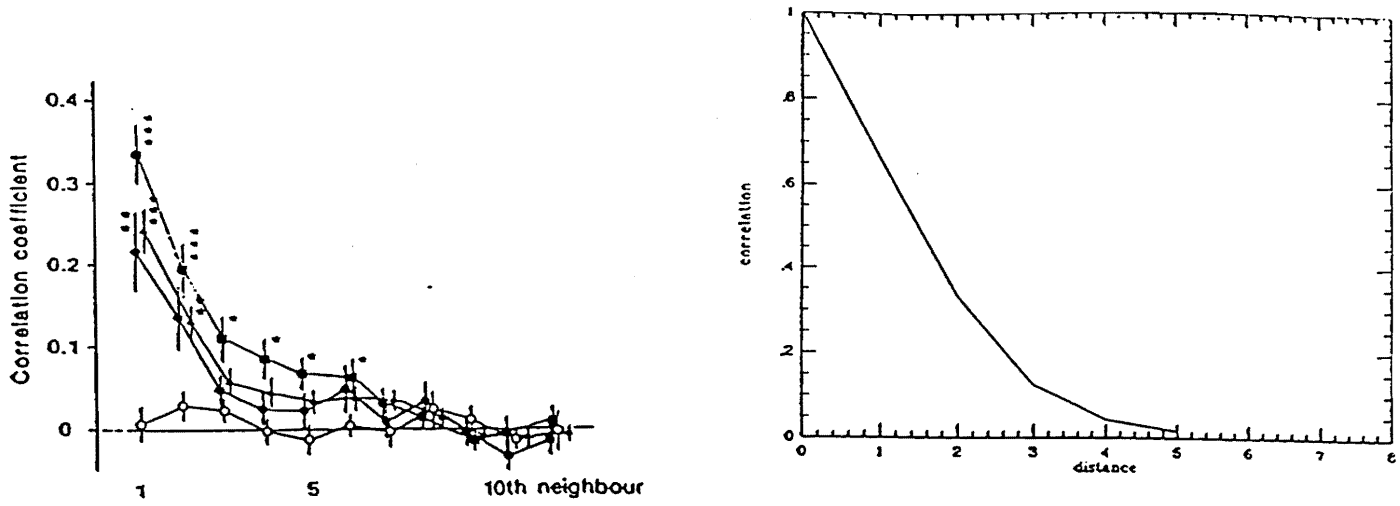

Figure 1: Correlations between attractors: left, delay activity population correlations (Miyashita); right, attractor correlations in the ± 1 model (Griniasti, Tsodyks, Amit).

process of training of the monkeys. The dynamics of the neurons, given the synaptic matrix, is:

$$S_i(t+\delta t) = \text{sign}[h_i(t+\delta t)], \tag{2}$$

where S_i are the 2-valued dynamical variables of a neuron and

$$h_i(t+\delta t) = \sum_{j=1}^{N} J_{ij} S_j(t). \tag{3}$$

h_i mimics the value of the post-synaptic potential, relative to the threshold, on neuron i. Note that for $a = 0$ we have the original Hopfield model, whose dynamics leads the network to attractors in which $S_i = \xi_i^\mu$, for large regions of stimuli around ξ^μ. What is represented in Fig 1b is the correlation between the attractors of the network defined by the above synaptic matrix, where different attractors are arrived at from different uncorrelated stimuli. The correlations are exactly the same for $0.5 < a < 1$.]

Learning, memory and cognitive psychology

The first observation about the model matrix is that it connects only contiguous patterns in the sequence. This is very significant, since in an attractor neural network the information about the previous stimulus is around in the active attractor when the next pattern is being presented. This information is there, but no more. Therefore, the fact that this information suffices in order to reproduce the full depth of the empirical correlations is both gratifying and surprising.

Note that this model, provoked directly by experiment, suggests a separation of diverse features identified in the Hopfield model and often taken to be an inseparable package.

- Attractor dynamics, which implies that a whole class of afferent stimuli leads to the same persistent reverberation, due to the extensive feed-back.
- Diversity through synaptic control, i.e. the fact that the synaptic structure, via learning, can produce a variety of attractors

- Auto-associativity, or the identification of the quality of the content addressable operation of the network with nearness of the internal representation in the attractor with the form of the stimulus used in learning.

What we find here is that experiment solicits the separation of the third feature from the first two, allowing for a much richer set of internal reverberation attractors. Those correlated attractors open a space for combining reverberations with computation.

As far as psycho-physics is concerned, attractor with this type of correlations can provide a building block in generating cognitive associations. Namely, we know from experiments on simulated networks (Rubin, Amit) that if a network is in a given attractor and a stimulus is presented which tries to move it to another attractor, the transition time is significantly shorter if the two attractors are correlated in their internal representations.

This is a typical effect of priming. But in the system under discussion we have more. We have an embryo of the formation process of priming: stimuli which are often presented contiguously in time are learned as correlated attractors and can then produce priming, as expressing or defining class partnership of the stimuli.

Granted monkeys may not be doing very highly cognitive functions and the class correlations described by the model are quite rudimentary. What is of significance is the appearance of a toy system comprised of a living brain, a cognitive psychology, an accompanying neuro-physiology, all tightly connected to a well understood quasi-realistic dynamical model.

What may be more important from a psychological point of view is that with the Miyashita monkey we find, I believe for the first time, a substrate for psychological construction which is irectly related to experiment. Ever more complex effects of cognitive phenomena may be interpretted in terms bridging between learning, stimulation and performance. Such constructions can then be tested by referring back to the monkeys, in an experimental context which is not metaphorical, provided language is not involved. To the extent that the dialectic between the non-linguistic psychological structures and the experiment is sufficiently confirmed, the extrapolation to linguistically expressed psychological mental events can turn out to be very fruitful. Some such beginnings can already be perceived.