Network 3 (1992) 379–384. Printed in the UK

LETTER TO THE EDITOR

Why the simplest notion of neocortex as an autoassociative memory would not work

Dominic O'Kane† and Alessandro Treves‡

† Theoretical Physics, University of Oxford, 1 Keble Rd, Oxford OX1 3NP, UK
‡ Department of Experimental Psychology, South Parks Road, Oxford OX1 3UD, UK

Received 23 December 1991

Abstract. We discuss the idea that long range cortico-cortical connections might be the substrate for an autoassociative memory mechanism, whereby features processed locally could be linked together over larger portions of neocortex. The simplest version of this idea is shown to be implausibly inadequate in terms of storage capacity: although up to a fraction of a bit could be stored on each synapse, the number of global activity patterns that could be stored and individually retrieved would scale not with the size of the network but, effectively, only with the number of modifiable connections per cell.

It is a widespread assumption (pervading, e.g., the discussions reported in [12]) that the substantial anatomical self-similarity of neocortical structure underlies a set of elementary operations that are carried out, on different incoming inputs but basically along the same lines, by different patches of cortex. One would like to grasp such (hypothetical) universal processing in simple conceptual terms. In this context, the system of local, or intrinsic, connections among pyramidal cells has often been thought, for example by Marr [9], to implement an autoassociative memory function, in the following manner. A (presumed) Hebb-like synaptic plasticity of these connections may enable them to store a set of local activity patterns. Later, afferent activity containing a fraction of the information associated with one such pattern—which would, by itself, elicit a distorted or partial version of the pattern—may trigger recurrent interactions through the local connections, resulting in the original activity pattern, or a very close version of it, with most of its information content, being relayed on for the next stage of processing. This memory retrieval may of course be compatible with other functions being performed concurrently, even by the same set of local connections; at the same time, this could be a way to describe in abstract terms a ubiquitous mechanism which would manifest itself in specialized forms, particular to the nature of the information being processed, in different cortical regions.

The number of synaptic connections which neocortical pyramidal cells receive from axons coming from the white matter is estimated to be, typically, of the same order as that of local excitatory connections from neighbouring pyramidal cells [1]. The great majority of those long-range connections originate in other neocortical areas. It is tempting, then, to extend the above hypothesis by considering the possibility that part of the long-range connectivity also operates as an autoassociative memory

0954-898X/92/040379+06$07.50

network. Possibly leaving aside feedforward and backprojections [13], such a role might be played by connections among cortical patches where information is processed simultaneously and in parallel. This would allow for activity patterns extending over large regions of cortex to be individually stored and retrieved.

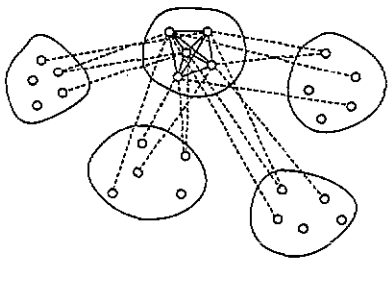
Similar views have inspired neurobiologists who tend to regard the neocortex as essentially a memory machine [4], and who have talked about functional units [5] consisting not of single cells but of local modules made up of many neighbouring cells [10]. Braitenberg [3] has suggested neglecting, as a first approximation, the complexity of local circuitry [8] as well as the specificity of cortico-cortical projections [6], and considering instead a simplified scheme, the 'skeleton' of neocortex, consisting only of its $\mathcal{N}$ pyramidal cells. If they were grouped into $\sqrt{\mathcal{N}}$ patches of $\sqrt{\mathcal{N}}$ cells each, any given patch could in principle be connected monosynaptically to any other patch by the axon of one cell.

For an autoassociative mechanism to operate successfully through connections that travel in the white matter, two obvious constraints must be satisfied: that the relative synapses be associatively modifiable, and that the longer conduction times be still compatible with the time scales for storage and retrieval. While no firm evidence exists on the first issue, and some amount of speculation is at present necessary to tackle the second, it is possible the both constraints be satisfied in neocortex. There is, however, at least a third question that has to be borne in mind, when judging how useful the notion of long-range (as opposed to local) autoassociative mechanisms is. The question is that of the efficiency with which large areas of neocortex would then operate as memory devices. We have addressed this issue by considering an appropriate formal model in the spirit of the 'skeleton' cortex, and calculating its capacity for storing activity patterns and information.

Let us consider a network of M modules, each containing N units. The short-range connectivity is complete, with each unit receiving inputs from all $N-1$ other units in its module, while the long-range connectivity is dilute and homogeneous, with each unit receiving inputs from L other units distributed at random among all other modules. Different modules process different aspects of an external input applied to the network, and this is modeled by assuming that D local *features* are stored on the local connections within each module, from where they may be individually retrieved with a local autoassociative mechanism. What are stored on long-range, intermodular connections, however, are P global activity patterns, each representing a combination of M features, one per module, as indicated in figure 1. If $P > D$, there will be P/D patterns which share the same feature in any particular module, but the full combination of features will be unique to any one pattern, and on average two given patterns will only share features in a fraction $1/D$ of the modules.

Activity patterns are stored on the reciprocated (symmetric) connections via a 'Hebbian' covariance learning rule, retrieval is taken to occur by means of attractor dynamics [2] with the information coded solely in the distribution of firing rates, and the remaining details of the model are taken as in [14, 15], where the applicability of capacity estimates to cortical situations has been discussed more extensively.

We present here the result of the capacity calculations, which will be described elsewhere [11]. The maximum number P_{max} of activity patterns that can be stored in, and retrieved from, the network, is in any case proportional to the total number of connections per unit $C = L + N - 1$, and depends on the sparseness a of the coding scheme in the usual way (cf [15]). We are interested in the way it depends on the fraction



module

pattern	1	2	3	4	5
1	**J**	**P**	**C**	**M**	**X**
2	I	Q	C	M	Z
3	K	P	B	L	Y
4	J	R	A	L	Z
5	K	R	B	N	X
6	I	Q	A	N	Y

Figure 1. The architecture of the network (top, indicating only the connections relative to one module) and the corresponding memory organization (middle). The table (bottom) gives a small scale example of how features could combine into patterns, when $\mu = 2$. Boldface letters denote one particular pattern being retrieved.

$$\gamma = \frac{L}{C} \tag{1}$$

of long-range connections, and on the ratio

$$\mu = \frac{P}{D} \tag{2}$$

measuring how many times a feature is repeated, locally, in different patterns. The 'square root' case of Braitenberg would correspond to $N = M = L = \sqrt{N}$ and $\gamma = 1/2$.

$P_{\max}$ turns out to be determined analytically by the expression

$$P = \frac{C}{A_3}\left[\frac{\mu(1-\gamma)}{(\gamma A_2 + \mu(1-\gamma)A_1)^2} + \frac{\gamma}{(\gamma + \mu(1-\gamma))^2 A_2^2}\right]^{-1}. \tag{3}$$

where the A symbols denote certain averages over the statistical distribution of the firing rates present in the activity patterns, and are defined in [15].

When $\mu = 1$, the organization of the memory is trivial, in the sense that each local feature pertains to only one global pattern (which is therefore uniquely identified, even looking at just one module); one is left with a mixed connectivity, partly long range and partly short range, and for the capacity we find results that interpolate exactly between the previously calculated capacities of networks with highly dilute and full connectivities.

With larger μ, we find that the usual contraint on the storage capacity remains local, expressed as a relation between D_{max} and N (the number of short-range connections per unit). This relation can be rewritten as

$$P_{max} \propto \mu(1-\gamma)C \tag{4}$$

and hence it would seem that, by increasing μ, P_{max} could be made large *ad libitum*. This is illusory, however, because of an insidious new phenomenon associated with the modular nature of the model: the appearance of the 'memory glass' state. In such a state (which bears some analogy, but is different from, the spin glass state of simple fully connected networks [2]) the network retrieves one of the very many possible *spurious combinations* of features, one that does not correspond to any of the stored global patterns. This state exists only for lower values of P than those limiting the existence of the genuine retrieval states, because its signal-to-noise ratio is lower (the signal being carried only by the fraction $(1-\gamma)$ of short-range connections).

When the memory glass state does exist, it takes up such a large basin of attraction, because of the very many spurious combinations it includes, that the network will tend to flow dynamically to it from most initial conditions. As μ becomes large, the interval of P values for which retrieval states still exist, but the threat of the memory glass does not, becomes narrow (figure 2), leaving one in the odd situation of having to fine tune the number of patterns stored in the network in order to ensure the proper retrieval of each one of them. In practice, we take this to mean that the high-μ region is not viable, and that as a result the pattern capacity is still effectively limited to the usual range proportional to N, with no special increase to be sought as a result of the organization of the patterns into features, no matter how many units ($\mathcal{N} \boxminus N \times M$) the network contains.

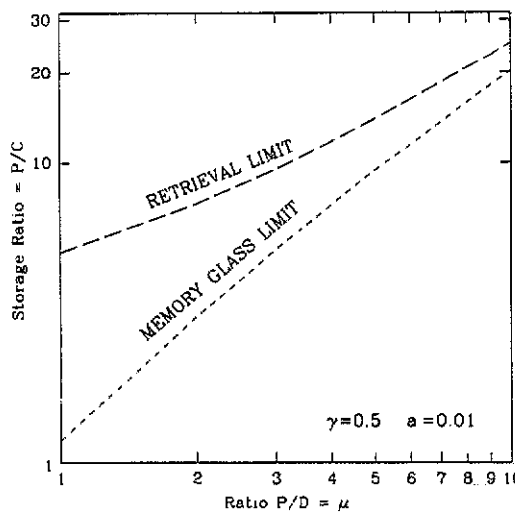


Figure 2. Capacity diagram, as a function of μ, for $\gamma = 1/2$ and $a = 0.01$. The upper line gives the limit P_{max} on the existence of retrieval states, whereas below the bottom line the memory glass state also exists and disrupts performance.

We have also computed the total information capacity of this network, and found it to remain in the usual [15] range of a fraction of a bit per synapse. One can

conclude, therefore, that our long-range autoassociative mechanism does not lead to under-using the available storage space, but rather to mis-managing it: the number of patterns stored cannot scale with the total size $\mathcal{N}$ of the network, and what scales with $\mathcal{N}$ is the information in bits contained in each pattern. Clearly, such a situation is untenable from the point of view of organizational efficiency: even without risking an arbitrary estimate [9] of the number of patterns a reasonable animal memory might store, one would like to guess that the extraordinary phylogenetic increase in neocortical size (*vis-a-vis* the much more modest increase in the number of synaptic contacts onto pyramidal cells) is to be ascribed to the need to store more memories, not just more complex ones.

What are the assumptions that determined our results, and that, if changed, would allow us to avoid the negative conclusion? One such assumption is neglecting (persistent) external inputs. A very strong external input (of the same order of the signal produced by local and long-range interactions) would stabilize memory patterns and destabilize the spurious combinations of the memory glass; but resorting to external inputs would be admitting the failure of long-range connections to perform their putative task, and it would also be against the notion that neocortex is close to a 'reflexive machine', working on its own output [3].

One possibility is that the assumption about the relevant neural code (the temporally coarse firing rate) be helplessly wrong. Abeles [1] has argued against neglecting the role that temporally fine phenomena such as synchronicity of firing might have in coding information, and the recent wave of experiments by Singer and his group [7] might be taken as lending support to this view.

It is also possible, on neuropsychological grounds, to point out the inadequacy of focusing solely on a memory ability for storing and retrieving discrete memory items [16]. It seems unlikely, however, that considering additional alternative forms of memory organization would by itself solve a problem associated, perhaps, with just one type of memory.

A much more conservative hypothesis is to note that maybe it is the assumption of a random long-range connectivity which produces the dismal performance. Indeed, such an assumption was only introduced as an interesting conceptual scheme, already known from the start to be at gross variance with the observed neuroanatomy. Possibly, capturing to just a slightly deeper level of detail the specificity of cortico-cortical connections might result in another organization of local activity patterns into memories, and solve the capacity problem. It is a challenge for theoreticians to produce a more sophisticated scheme, but with the same appealing simplicity as the random connectivity one, which would allow analysis and discussion of a model of the large scale properties of memory in neocortex.

References

[1] Abeles M 1991 *Corticonics* (Cambridge: Cambridge University Press)
[2] Amit D J 1989 *Modeling Brain Function* (Cambridge: Cambridge University Press)
[3] Braitenberg V 1978 Cortical architectonics: general and areal *Architectonics of the Cerebral Cortex* ed M A B Brazier and H Petsche (New York: Raven Press) pp 443–65
[4] Braitenberg V and Schütz A 1991 *Anatomy of the Cortex: Statistics and Geometry* (Berlin: Springer)
[5] Eccles J C 1984 The cerebral cortex, a theory of its operation *Cerebral Cortex* ed A Peters and E G Jones, vol 1 (New York: Plenum) pp 1–36
[6] Goldman-Rakic P S 1988 Changing concepts of cortical connectivity: parallel distributed cortical networks *Neurobiology of Neocortex* ed P Rakic and W Singer (New York: Wiley) pp 177–202

[7] Gray C M, König P, Engel A K and Singer W 1989 Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties *Nature* **338** 334–7

[8] E G Jones 1988 What are the local circuits? *Neurobiology of Neocortex* ed P Rakic and W Singer (New York: Wiley) pp 137–52

[9] Marr D 1970 A theory for cerebral neocortex *Proc. R. Soc.* B **176** 161–234; 1971 Simple memory: a theory for archicortex *Phil. Trans. R. Soc.* B **262** 23–81

[10] Mountcastle V B 1979 An organizing principle for cerebral function: the unit module and the distributed system *The Neurosciences: Fourth Study Program* ed F O Schmitt and F G Worden (Cambridge, MA: MIT Press) pp 21–42

[11] O'Kane D and Treves A 1992 Short- and long-range connections in associative memory *J. Phys. A: Math. Gen.* **25** 5055–69

[12] Rakic P and Singer W (ed) 1988 *Neurobiology of Neocortex* (New York: Wiley)

[13] Rolls E T 1989 The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampus *The computing neuron* ed R Durbin, C Miall and G Mitchison (Reading, MA: Addison Wesley) pp 125–59

[14] Treves A 1990 Graded-response neurons and information encodings in autoassociative memories *Phys. Rev.* A **42** 2418–30

[15] Treves A and Rolls E T 1991 What determines the capacity of autoassociative memories in the brain? *Network* **2** 371–97

[16] Weiskrantz L 1990 Problems of learning and memory: one or multiple memory systems? *Phil. Trans. R. Soc.* **329** 99–108