Neural networks in neuroscience: a brief overview

Samuel Johnson[1]

Instituto Carlos I de Física Teórica y Computacional, and Dpto. de Electromagnetismo y Física de la Materia, Universidad de Granada, 18071 Granada, Spain.

Ever since the publication of Santiago Ramón y Cajal's drawings of neurons – in his words, those "mysterious butterflies of the soul" – it has been clear that the nervous system is composed of a large number of such cells connected to one another to form a network. Long axons, ending in terminals which form synapses to the dendrites which branch out from neighbouring neurons, transmit bursts of electric current and enable neurons somehow to cooperate and yield the astonishing emergent phenomenon known as thought. However, the concept of a neural network (as understood in theoretical and computational neuroscience) was partly inspired by the mathematical model first studied by Ernst Ising in the early 1920s with a view to understanding magnets. It was known that the spin of electrons conferred a magnetic moment to individual atoms, but it wasn't clear how a very many such spins could self-organise into a large body producing a net magnetic field. By considering an infinite set of entities (spins) with possible values plus or minus one which, when placed at the nodes of a lattice, interact by trying to align themselves with their neighbours, and a temperature parameter to govern the extent of random fluctuations, it was eventually shown that, below a certain critical temperature (in two or more dimensions), symmetry is spontaneously broken and most of the spins end up aligned. This *ferromagnetic* solution comes about and is then maintained because it has a lower energy than any other configuration of spins.

In an effort altogether independent of statistical mechanics but which related Pavlovian conditioning experiments with cellular plasticity, the psychologist Donald Hebb conjectured, in 1949, the existence of some biological mechanism which would lead to neurons which repeatedly fired (i.e., let off bursts of current, or *action potentials*) together becoming more strongly coupled [1]. This idea of *Hebbian learning* was used by John Hopfield in 1982 for his model of a neural network [2]. The situation considered was basically an Ising model in which the spins were substituted for "binary neurons" – elements which could either fire or not at a given moment. Because neurons fire when the total electric current received from their neighbours depolarises their membranes above a characteristic threshold, a given neuron is more likely to fire if its neighbours are firing, resulting in a dynamics similar to that of Ising's model atoms. However, a key difference in the Hopfield model is that the interaction strength of the connections between neurons – the synaptic weights – can be varied so as to store information via Hebbian learning. Imagine, for the sake of illustration, that the set of neurons is laid out on a plane, the ones with a positive activity (i.e., firing) coloured black and the others white. In this way, we can represent a black and white pattern as a particular configuration of activities, just as a computer might store an image as a

[1] samuel@onsager.ugr.es

bitmap. Taking each pair of connected neurons and either reinforcing their mutual interaction strength if the activities are alike, or reducing it when they are opposite, the minimum energy configurations will no longer correspond to all black or all white, as in the Ising model, but rather to a configuration which coincides with the stored pattern – or to its negative. Now, if the temperature is below the critical value, the system will always self-organise into either the pattern or the anti-pattern whatever the initial configuration. Mathematically, these states have become *attractors* of the dynamics.

The mechanism described is known as associative memory. Importantly, the network isn't limited to learning only one pattern. We can take a whole set of different patterns and, for each one, perform the same operation of adding or subtracting a little strength to each synapse, with the result that each pattern (and each anti-pattern) becomes an attractor. The system will, in general, evolve towards whichever of the patterns most resembles the configuration we place it in initially. For instance, if we store a set of photos of various people and then "show" the network a different picture of one of the same subjects, it can retrieve the correct identity. Not only is this mechanism used nowadays in technology capable of performing tasks such as pattern discrimination and classification, but it is widely considered to underlie our own capacity for learning and recalling information.

Not long ago, *in vivo* experiments finally established that learning is indeed related to the phenomenon of long term potentiation (LTP), by which synapses between neurons that fire nearly simultaneously gradually increase their conductance [3]. Even more recent evidence from micro-arrays placed on the cortex has supported the idea that certain patterns of activity code for particular memories. The neural network models studied nowadays generally include more realistic dynamics both for the neurons and for the synapses, taking into account a variety of cellular and subcellular processes. For example, the fact that the conductance of synapses in reality depends on their workload has been found to enable a network to switch from one pattern to another – either spontaneously or as a reaction to sensory stimuli – providing a means for the performance of dynamic tasks [4,5]; this result also seems to agree well with physiological data [6]. In fact, there is mounting theoretical [7] and experimental [8] evidence that the brain somehow maintains itself close to a boundary – called, in physics, a critical point – between an ordered and a chaotic regime.

That these models should actually reflect, albeit in an enormously simplified way, what actually goes on in our brains tends to fit in quite well with intuitive expectations. For instance, from this point of view the way in which the recollection of a particular detail often evokes, almost instantly, a whole scenario of sensations and emotions makes sense, since these concepts will have been stored in some way as the same pattern. Also, the fact that new memories are recorded in synapses which were already being used to store previous information would seem to explain why memories tend to fade slowly away with time, yet can still be recalled, at least to some extent, when a particular thought in some sense overlaps with (reminds one of) one of them. When this happens, the old memory springs to mind and, if held there long enough, can be refreshed via

LTP – although interaction with other patterns or with current stimuli may well modify the refreshed information. Similarly, previous information influences the storing of new memories, leading to the well known fact that we tend to “see” things the way we expect them to be. This brings us to the question of how many patterns can be stored in a given network. In the Hopfield model, if each neuron is connected to all the rest – a mathematically convenient situation which usually has little to do with reality – then the maximum storage capacity is about 0.14 completely distinct patterns per neuron. However, the *topology* [9] of a network – how the synapses are placed among the neurons – and the dynamics of the synapses [10] seem to have a strong bearing on this as well as other features of neural networks.

Scientists currently developing and studying this kind of models can classify the tasks lying ahead of them in two broad categories: on the one hand, to continue to address features of brain function which are as yet little understood, such as short-tem memory, information processing, emotional tagging, or the concepts of *consciousness* and *understanding,* to name but the first few that spring to mind; and on the other, to convince those researchers from other areas – especially biologists and psychologists, but also doctors, artificial intelligence workers, linguists and many others – which remain sceptical or are simply unaware of such theoretical results that these must be taken into account if progress in neuroscience is to be achieved.

References

1. D.O. Hebb, *The Organization of Behavior,* Wiley, New York, 1949.
2. J.J. Hopfield, *Neural networks and physical systems with emergent collective computational capabilities,* PNAS **79**, 2554 (1982).
3. A. Gruart, M.D. Muñoz, and J.M. Delgado-García, *Involvement of the CA3-CA1 synapse in the acquisition of associative learning in behaving mice,* J. Neurosci. **26,** 1077 (2006).
4. J.M. Cortés, J.J. Torres, J. Marro, P.L. Garrido, and H.J. Kappen, *Effects of fast presynaptic noise in attractor neural networks,* Neural Comp., **18**, 614(2006).
5. H.D. Holcman and M. Tsodyks, *The emergence of Up and Down states in cortical networks,* PLoS Comput. Biol. **2** e23 (2006).
6. H. Korn and P. Faure, *Is there chaos in the brain? II. Experimental evidence and related models,* C. R. Biol. **326**, 787 (2003).
7. S. Johnson, J. Marro, and J.J. Torres*, Functional optimization in complex excitable networks,* EPL **83**, 46006 (2008).
8. V.M. Eguíluz, D.R. Chialvo, G.A. Cecchi, M. Baliki, and A.V. Apkarian, *Scale-free brain functional networks*, Phys. Rev. Lett. **94**, 018102 (2005).
9. J.J. Torres, M.A. Muñoz, J. Marro, and P.L. Garrido, *Influence of topology on the performance of a neural network,* Neurocomputing **229**, 58 (2004).
10. J. F. Mejias and J. J. Torres, *Maximum memory capacity on neural networks with short-term synaptic depression and facilitation*, Neural Comput., **21** (3), 851 (2009).