BOOK AND NEW MEDIA REVIEWS

WHO OPENED PANDORA'S BOX?

Review of *Computational Neuroscience of Vision* by Edmund T. Rolls and Gustavo Deco. ISBN 0-19-852488-9 (pbk), Oxford: Oxford University Press, 2002, 564 pages, Price US $55.00; 38.46; UK £26.95

The cover of Rolls and Deco's book *Computational Neuroscience of Vision* depicts a painting of J.W. Waterhouse surrounded by a diagram of the visual system modelled by neural networks. The picture shows Pandora opening the infamous box. This promotes a common misconception of who actually opened Pandora's box.

According to Greek mythology Pandora was fashioned from clay at the request of Zeus. She was blessed with every gift the gods could grant and Zeus endowed her with a box scheming to destroy Prometheus' creation of man. Realising that Prometheus would be too wise to accept the box as a dowry Zeus conducted Pandora to his less cautious brother Epimetheus. Pandora was so beautiful and irresistible Epimetheus simply could not refuse. He opened the box thereby unleashing all the evils and diseases to afflict human life ever since. Only Hope lingered at the bottom of the box to console mankind in his troubles.

The desire to open the black box, that is to understand the functioning of the human brain, has been at the heart of neuroscience. Rolls and Deco seek understanding through neural computation but did they manage to open the black box? If so were they less cautious than other neuroscientists even releasing bad spirits that will haunt computational neuroscience for eternity? In the following we will give a brief outline of the book before we critically assess their approach.

The title of the book suggests a broad introduction to computational neuroscience of vision but the early stages of encoding are only touched upon: the optics of the eye, colour, contrast, spatial and temporal filtering, disparity and motion are mentioned in passing. These topics are widely regarded as the foundations of vision (e.g., Palmer, 1999; Wandell, 1995) whereas object recognition and visual attention are at the high-end of visual processing. In fact Rolls and Deco are mainly concerned with high-level performances that carry a considerable cognitive load.

The first six chapters describe physiological aspects of cell activation and the anatomical structure and function of many of the cortical areas involved in visual processing, including visual areas in the temporal lobe where objects are represented. They introduce important physiological constraints and hypotheses that are exploited later. Chapter 7 gives a comprehensive account of neural network models together with some of their basic properties. Readers who are familiar with Rolls and Treves (1998) can quickly browse through this chapter to recall different architectures. Their effort to explain the different architectures has to be applauded. Layout and figures are generally of high standard but some equations, the Hebb learning rule and the activation rule in particular make too many appearances throughout the book. Redundancy can be helpful but a more stringent organisation of the theoretical backbone would be preferred. Chapters 8 to 12 describe applications of various network architectures to the problem of visual object recognition and visual attention. In object recognition, the visual system has to establish a representation of objects that allows recognition independent of location, size, orientation, contrast, illumination etc. In Chapter 8 VisNet, a feature-hierarchy network model, is introduced which consists of four hierarchically connected competitive networks. Representation of objects is achieved from spatial convergence and feature integration by forward connections between hierarchically organised layers, and from redundancy removal within each layer. After training of each layer VisNet exhibited translation and view invariance in a face-recognition task. In a variant of the model the network was extended to a recurrent attractor network that is influenced by top-down bias. In Chapter 9 it is illustrated how the interaction between object and spatial processing implemented by back-projections from the ventral and dorsal stream to an early topologically organised visual area can account for many properties of visual attention. Effects of attention on single cell and fMRI recordings as well as psychophysical performance are simulated. In Chapter 10 the model is extended and aspects of visual search and attention are simulated. The system essentially works in parallel but due to the different latencies of its dynamics two experimentally observed modes of visual attention can be modelled: Serial focal attention and parallel spread of attention over space. The binding problem in conjunction search requires additional feature maps (i.e., size and colour). Confing the model in different ways a variety of dysfunctions associated with visual neglect are simulated in Chapter 11. In Chapter 12 the focus is on outputs of IT via perihinal and entorhinal cortex to the hippocampus as well as the orbitofrontal cortex and amygdala. These structures are used to model

phenomena of short- and long-term memory, emotion, reward and punishment, visual search and attention. The achievements of various implementations in the book are summarized in the last chapter.

Many of the goals of machine intelligence are accomplished within biological nervous systems, by using strategies, architectures and hardware (or rather 'wetware') radically different from the common serial computer (von Neumann architecture). For example, in living neural systems there are no formal or numerical representations, communication media are stochastic, events are asynchronous, components are unreliable and widely distributed, connectivity does not obey precise blueprints, and processing speeds are much slower. Yet the performance of natural systems in real-time tasks entailing perception, learning, and motor control, in complex environments, remains unrivalled. A central goal of computational neuroscience is therefore to understand how this is possible, and to exploit and incorporate the underlying computational principles within new artificial systems.

Originally neural networks were relatively simple input-output devices with two layers of interconnected neurons. The Perceptron, Adeline, the Boltzmann machine, and the Hopfield model are prominent examples. At the time they competed with serial computers showing some striking advantages. It is well known that neural networks work in parallel with distributed representations, they learn from past experience, can handle noise and incomplete data, and once trained neural networks respond to new input almost instantly. Also the same neural software can be applied to a number of different problems (see also, Aleksander, 1989; Bishop, 1995; Haykin, 1994).

As a consequence of their universality and omnipotence the input to neural networks is crucial. This is not at all trivial for visual input. It needs to be decided whether images should contain colour, disparity, and motion for example, and how they are standardised or sampled along various dimensions such as position, size, luminance, contrast, illumination, viewpoint, etc.

According to folklore in computer science the Pentagon commissioned in the 80s a costly research program to build a system that could quickly detect whether a natural image contained a tank or not. A set of images with and without tanks was split into half and a neural network was trained on half of the images. After training the system performed perfectly on the first half. When the untrained second set of images was employed the system still performed perfectly. Apparently it could detect features of tanks in natural scenes. Only when the system was put to the test with a new set of images it failed miserably. What had happened? The original set of images systematically differed in their illumination with tanks photographed on a cloudy day while the images without tanks on a sunny day. The network had classified the images accordingly. Relevant features were superseded by the irrelevant illumination cue that did not generalise.

This anecdotal evidence illustrates one of the problems with neural networks. A trained network with more than a few dozen neurons is difficult to analyze and understand. A neural network, especially in a complex architecture and trained with complex input, cannot explain its output. It essentially remains a black box. One of the dangers of using multi-layer architectures and unsupervised learning is this lack of insight. The system may perform perfectly well with a given training set but the experimenter remains ignorant of the discriminating features. Problems only transpire when generalisation does not occur with novel input (e.g., Hecht-Nielson, 1991; Hertz et al., 1991).

The problems associated with computational models of vision are comparable to the general problems encountered in Artificial Intelligence: Specific implementations lack the general capabilities of the human visual system. As a consequence the field has become more specialised, focussing on separate modules and functions with little emphasis on an integrated framework. In this respect Rolls and Deco present a refreshing alternative. They employ neural networks as modules in various architectures to model aspects of object recognition, visual attention and beyond.

There are, however, disadvantages inherent to neural networks that remain virulent in more complex architectures. The 'curse of dimensionality' for example refers to the problem that the quantity of training trials can grow exponentially with the number of input variables. A related problem is overfitting. Due to the substantial amount of free parameters the system may not only learn relevant features but noise as well. The consequence is suboptimal performance as the system looks not only for relevant features but also for irrelevant noise patterns in novel images.

The problem of dimensionality and overfitting has implications for studies reported in Chapter 8. Discrimination performance is illustrated for relatively few objects: 3 line stimuli in 9 locations, 7 faces in 9 locations, 3 faces from 7 different views. VisNet has 1024 output units in the fourth layer alone to establish an object representation. It would be interesting to see how the system learns and performs as it approaches at least half of its theoretical capacity. The authors are suspiciously unconcerned about training trials and phenomena linked to perceptual learning.

When VisNet was trained with 7 face stimuli presented on a blank background and tested on a cluttered background discrimination performance

was good. Performance was poor however when the faces were first learned on a cluttered background and then tested on a blank background (Stringer and Rolls, 2000). The authors concede that segmentation through depth, motion and colour is required as well as attention to perform well in this task and refer to later chapters. But evidence for recognition in cluttered scenes remains scarce (see their Fig. 9.8 on page 342).

Various network architectures are suggested that can model high-level functions of object recognition and attention (hierarchical feed-forward competitive networks, recurrent attractor networks, networks with top-down influence, inhibitory pools, back-projections, etc). This brings about additional complexity and degrees of freedom. It is not surprising that these networks can perform with very few objects after suitable training but what can be generalised to novel input. How can we falsify a particular architecture? The authors occasionally derive predictions and conduct simulations but mostly they illustrate the capabilities of their network models without thoroughly testing their systems.

The assumption that processing is confined to spatial features is only a first step. In early stages of processing the human visual system uses a variety of dedicated receptors and filters. Different cell types in LGN and V1 are known to extract not only spatial information and colour but also temporal aspects (DeAngelis et al.,1995). It has been suggested that in addition to the 'what' and 'where' stream the medial temporal cortex (MT/MST) constitutes a 'when' stream because it mediates between dorsal and ventral stream (e.g., Kourtzi et al., 2001). The representation of moving objects with features defined in space and time is neglected in their models. Representation of moving objects is limited to snapshots contiguous in time. It seems reasonable to assume that motion as well as depth processing contributes to translation and view-invariance in object recognition as the visual system is capable of encoding and storing dynamic information of complex moving objects such as biological motion.

The book is inspiring and well written. Without a doubt the cross talk between neuroscience and computational modelling will continue and advances in neurophysiology and the design of neural networks will help to develop a more unified architecture. Whether piecing together modules of neural networks is sufficient to create an integrated model of the visual system remains questionable. It is an ambitious goal and worthwhile to pursue but its limitations closely resonate the deep problems encountered in artificial intelligence.

We do believe that Rolls and Deco opened the black box to some extent. A number of controversial ideas have been released which are likely to stay with us for some time but not for eternity. Even though network models can be something of a black box themselves, the combination of parallel and hierarchical processing is probably the only way to overcome the physiological constraints of an integrated visual system. Rolls and Deco successfully demonstrate how versatile neural network models can be and how many puzzling phenomena of object recognition and visual attention could be solved in a general framework. This does not automatically secure Rolls and Deco a place in the pantheon of immortal neuroscientists but there is always Hope at the bottom of the black box.

Martin Lages and Alexander Dolia

References

ALEKSANDER I. *Neural computing architectures*. North Oxford Academic Press, 1989.

BISHOP CM. *Neural networks for pattern recognition*. Oxford: Clarendon Press, 1995.

DEANGELIS GC, OHZAWA I and FREEMAN RD. Receptive-field dynamics in the central visual pathways. *Trends in Neuroscience*, *18*: 451-458, 1995.

HAYKIN S. *Neural networks: A comprehensive foundation*. Macmillan, 1994.

HECHT-NIELSON R. *Neurocomputing*. Reading MA: Addison-Wesley, 1991.

HERTZ JA, KROGH A and PALMER RG. *Introduction to the theory of neural computation*. Wokingham: Addison Wesley, 1991.

KOURTZI Z, BÜLTHOFF HH, ERB M and GRODD W. Object-selective responses in the human motion area MT/MST, *Nature Neuroscience*, *5*: 17-18, 2002.

PALMER SE. *Vision Science. Photons to phenomenology*. Cambridge MA: MIT Press, 1999.

ROLLS ET and TREVES A. *Neural networks and brain function*. Oxford: Oxford University Press, 1998.

STRINGER SM and ROLLS ET. Position invariant recognition in the visual system with cluttered environments, *Neural Networks*, *13*: 305-315, 2000.

WANDELL BA. *Foundations of vision*. Sunderland MA: Sinauer, 1995.

Martin Lages, Department of Psychology, University of Glasgow, 58 Hillhead Street, Glasgow
G12 8QB. Email: m.lages@psy.gla.ac.uk;
Webpage: http://www.psy.gla.ac.uk/index.php?section=staff&id=ML001
Alexander Dolia, School of Electronics and Computer Science, University of Southampton
Southampton, SO17 1BJ, England. Email: ad@ecs.soton.ac.uk;
Webpage: http://www.ecs.soton.ac.uk/info/people/ad