Neuroscience: New Insights for AI?

Tomaso Poggio

McGovern Institute for Brain Research
Center for Biological and Computational Learning
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Abstract. Understanding the processing of information in our cortex is a significant part of understanding how the brain works and of understanding intelligence itself, arguably one of the greatest problems in science today. In particular, our visual abilities are computationally amazing and we are still far from imitating them with computers. Thus, visual cortex may well be a good proxy for the rest of the cortex and indeed for intelligence itself. But despite enormous progress in the physiology and anatomy of the visual cortex, our understanding of the underlying computations remains fragmentary. This position paper is based on the very recent, surprising realization that we may be on the verge of developing an initial quantitative theory of visual cortex, faithful to known physiology and able to mimic human performance in difficult recognition tasks, outperforming current computer vision systems. The proof of principle was provided by a preliminary model that, spanning several levels from biophysics to circuitry to the highest system level, describes information processing in the feedforward pathway of the ventral stream of primate visual cortex. The thesis of this paper is that – finally – neurally plausible computational models are beginning to provide powerful new insights into the key problem of how the brain works, and how to implement learning and intelligence in machines.

I have always believed that theoretical results from information theory, theory of computation, and learning theory will play an important role in our understanding of how the brain processes information and how intelligent behavior arises from a large number of neurons. At the same time, I felt that the gap between computer science and neuroscience was still too large for establishing a direct connection. Until a few months ago, I always tried to keep separate the projects in my lab focusing on computer vision, i.e. developing engineered systems for image recognition, from the projects focused on the functions of visual cortex.

A few months ago, for the first time in my career, my perspective changed in a dramatic way. The turning point was a surprising discovery: a preliminary model implementing the theory of visual cortex on which we have been working for the last five years, in close cooperation with a number of anatomical and electrophysiological labs, turned out to perform as well or better than the best engineering systems and as well as humans on difficult recognition tasks involving

N. Zhong et al. (Eds.): WImBI 2006, LNAI 4845, pp. 32–35, 2007.

natural, complex images. In my mind this meant that we may be closer to a basic understanding of how visual cortex recognizes objects and scenes than I ever thought possible. It also means that the AI community should follow this kind of developments in neuroscience quite closely. Let me first describe the problem, its importance, and then the approach that I propose.

Specific problem: The human visual system rapidly and effortlessly recognizes a large number of diverse objects in cluttered, natural scenes. In particular, it can easily categorize images or parts of them, for instance faces, and identify a specific one. Despite the ease with which we see, visual recognition – one of the key issues addressed in computer vision – has remained quite difficult for computers and is indeed widely acknowledged to be a very difficult computational problem. Object recognition in primate cortex is thought to be mediated by the ventral visual pathway running from primary visual cortex, V1, over extrastriate visual areas V2 and V4 to inferotemporal cortex, IT. IT in turn is a major source of input to PFC involved in linking perception to memory and action. Over the last decade, several physiological studies in non-human primates have established a core of basic facts about cortical mechanisms of recognition that seem to be widely accepted and that confirm and refine older data from neuropsychology. Given the wealth of physiological and behavioral data do we understand how visual recognition is done? Can we develop a theory leading to computer models capable of processing images as visual cortex does?

Why developing a theory is both difficult and important: After the breakthrough recordings in V1 by Hubel and Wiesel there has been a noticeable dearth of comprehensive theories attempting to explain the function and the architecture of visual cortex beyond V1. The reason of course is that a comprehensive theory is highly constrained by many different data from anatomy and physiology at different stages of the ventral stream and by the requirement of matching human performance in complex visual tasks such as object recognition. Thus, developing a consistent, quantitative theory is difficult. However, it would be extremely useful. Even a partial understanding of visual cortex is likely to provide powerful insights in how other parts of cortex work. Finally, theoretical foundations would be of key importance for the AI community because ultimately we want to understand the information processing involved in seeing and be able to replicate it in machines.

Preliminary results: One of the first models of visual object recognition, Fukushima's Neocognitron (Fukushima, 1980), followed the basic Hubel and Wiesel hierarchy (Hubel and Wiesel, 1968) in a computer vision system. Building upon several conceptual proposals (Perrett and Oram, 1993; Wallis and Rolls, 1997; Mel, 1997), we developed (Riesenhuber and Poggio, 1999; Serre et al., 2002; Giese and Poggio, 2003) a similar computational model. The present theory (Serre et al., 2005) has evolved over the last 6 years from that initial model. The theory is the outcome of computer simulations, trying to quantitatively account for a host of recent anatomical and physiological data. It is mainly the result of collaborations and interactions with several neuroscience experimental

labs (N. Logothetis in the early years and now D. Ferster, E. Miller, J. DiCarlo, C. Koch, I. Lampl, W. Freiwald, M. Livingstone, E. Connor). The architecture of the model resulting from the theory is shown in Fig. 1. It is qualitatively and quantitatively consistent with (and in some cases actually predicts) several properties of cells in V1 (Lampl et al., 2004), V2, V4 (Gawne and Martin, 2002) and IT (Logothetis et al, 1995; Hung et al., 2005) as well as fMRI and psychophysical data (Riesenhuber et al., 2004). The present theory bridges several levels of understanding, from computation and psychophysics to system physiology and anatomy, to the level of specific microcircuits and biophysical properties. The key extension with respect to the original model by Riesenhuber and Poggio is an unsupervised learning of the tuning of each unit at the S2, S2b and S3 levels (possibly corresponding to V4 and PIT, see Fig. 1) on a set of natural images unrelated to the task. In the present model, units (of the simple type) become tuned to the neural activity induced by natural images within their receptive field.

The most interesting, new result is that the model (see Fig. 1) outperforms the best computer vision systems on several different recognition tasks on real-world natural images. In fact, this is perhaps the first time that a model of cortex does as well as humans on a natural image recognition task. Even more surprisingly, the model mimics human performance when tested for rapid categorization without eye movements. The full theory and the results above are still unpublished (apart from a technical report titled, "A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex" by Serre, Kouh, Cadieu, Knoblich, Kreiman and Poggio, 2005).

I should emphasize that this is still far from solving the problem of vision. First, vision is more than object recognition and the visual cortex is more than the ventral stream. Second, the model in its present form cannot account for normal, everyday vision which involves eye movements and complex attentional top-down effects which must be mediated by higher brain centers and the extensive anatomical backprojections found throughout visual cortex. However, this theory may account for the immediate recognition of single pictures – a task humans can perform very well.

The open question – beyond establishing the basic aspect of the feedforward model – is whether it can be extended in the next few years to become a full theory of normal vision. The obvious approach involves physiology, psychophysics, fMRI and AI. In this project graduate students from computer science would work along with neuroscience students trying to solve simultaneously the problem of how visual cortex works and how to build a machine that sees.

In summary, the time may have come for AI to learn from serious neuroscience.

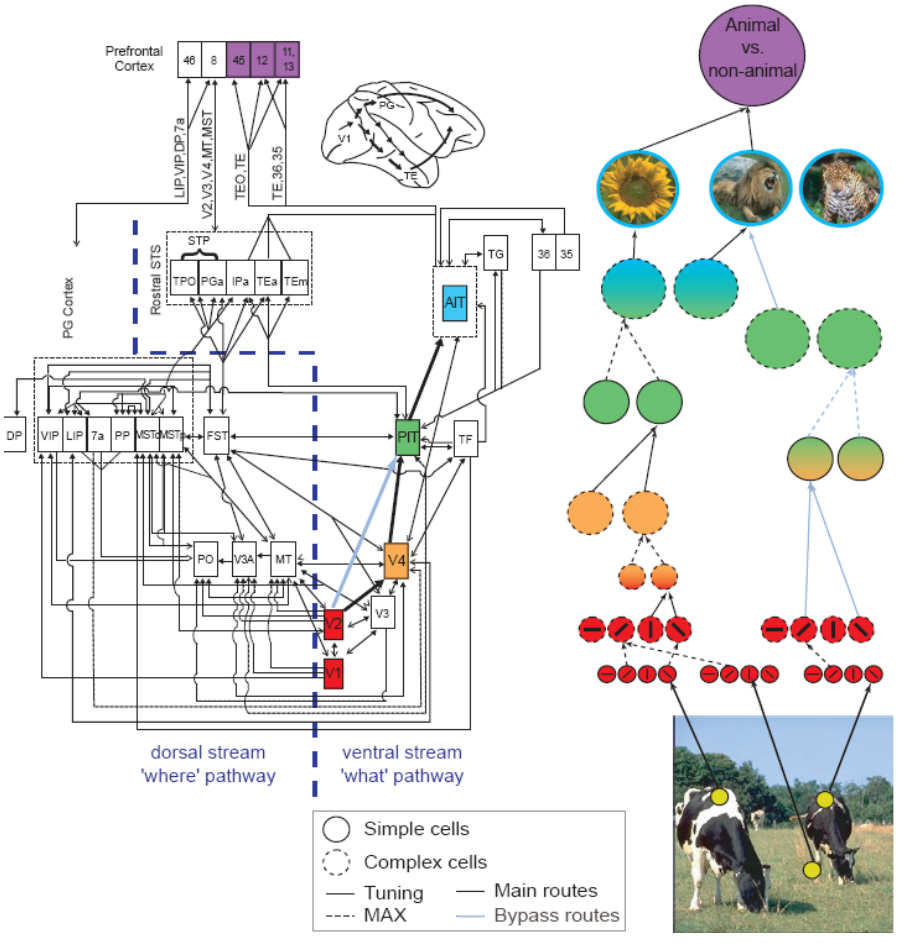

Fig. 1. The basic architecture of the model of the ventral stream (right). The figure provides a tentative mapping to the structural primitives of the ventral stream in the primate visual system (left). The theory assumes that one of the main functions of the ventral stream is to achieve a trade-off between selectivity and invariance. There are two basic operations iterated thoughout the hierarchy. Stages of "simple" (S) units with Gaussian-like tuning (plain circles and arrows), are interleaved with layers of "complex" (C) units (dotted circles and arrows), which perform a max operation on their inputs and provide invariance to position and scale. Developmental-like unsupervised learning, on a set of natural images, determines the tuning of the simple units in the S2 and S3 layers (corresponding to V4 and PIT, respectively). Learning of the synaptic weights from S4 to the top classification units is the only task-dependent, supervised learning stage in this architecture. The total number of units in the model is in the order of 2^7. Colors indicate the correspondence between model layers and cortical areas. The table on the right provides a summary of the main properties of the units at the different levels of the model. The diagram on the left is modified from Van Essen and Ungerleider (with permission).