Immediate recognition and feedforward models: what is next?

t.poggio, CBCL
Ventral stream in visual cortex: V1 to IT

Can the basic Hubel and Wiesel simple-complex architecture (hierarchical, feedforward only) be extended up to IT?

As discussed in SUnS 06 by Serre, Thorpe, Treisman… the answer seems (surprisingly) to be yes (with caveats)
Feedforward models for object/scene recognition: background

- Computer vision: *feedforward hierarchical models of object recognition* (LeCun, SungPoggio, Kanade, Viola, Amit, Ullman…)

- Neurobiological “models” (Hubel & Wiesel, 1959; Fukushima, 1980; Oram & Perrett, 1993, Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999; Thorpe; Mel; Koerner…)
Most detailed feedforward model – as far as I know – of the ventral stream (for object recognition)

Thomas Serre, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, Gabriel Kreiman, A. Oliva and Tomaso Poggio, 2005
2 biophysically plausible operations (by small circuits of spiking neurons):

1. \(~\text{Max}~\) (via shunting inhibition)
2. \(~\text{Gaussian}~\) (~dissive normalization)

Predicts/matches properties of cortical neurons in V4, IT, PFC
(Serre, Kouh, Cadieu, Knoblich, Kreiman, Poggio, 2005)
One example in V4

C2 units

V4 neurons

[Reynolds, Desimone et al., 1999]
- **Task-specific circuits** (from IT to PFC)
  - Supervised learning: ~ Gaussian RBF/linear classifier

- **Generic dictionary of reusable shape components** (from V1 to IT)
  - Unsupervised learning (from ~10,000 natural images) during a developmental-like stage
Partial list of detailed ‘predictions’ of physiology and psychophysics data

- MAX in V1 (Lampl et al, 2004) and V4 (Gawne et al, 2002)
- Tow-spot reverse correlation in V1 (Livingstone and Conway, 2003; Serre et al, 2005)
- Tuning for boundary conformation (Pasupathy & Connor, 2001) in V4
- Tuning for gratings in V4 (Gallant et al, 1996; Serre et al, 2005)
- Tuning for two-bar stimuli in V4 (Reynolds et al, 1999; Serre et al, 2005)
- Tuning to Cartesian and non-Cartesian gratings in V4 (Serre et al, 2005)
- Two-spot interaction in V4 (Freiwald et al, 2005; Cadieu, 2005)
- Tuning and invariance properties in AIT (Logothetis et al, 1995)
- Average “average effect” in IT (Zoccolan, Cox & DiCarlo, 2005)
- IT read out data (Hung et al, 2005)
- Trade-off of selectivity and invariance in IT (Zoccolan, Kouh, DiCarlo, 2007)
- Face inversion effect (Riesenhuber, Sinha et al, 2004)
- Rapid categorization (Serre et al., 2005)
As reported by Serre last year, the model seems to do well compared with state-of-the-art computer vision systems on natural images (StreetScenes, Caltech 101 etc.; see Bileschi thesis) ...

...and seems to account for rapid categorization of complex natural images by humans

(Serre, Oliva, Poggio, PNAS in press) . . .
Rapid categorization task (with mask to test feedforward model)

Animal present or not?

30 ms ISI

(Potter, Thorpe et al, 1996; Van Rullen & Koch, 2003; Bacon–Mace et al, 2005; Oliva & Torralba, in press)
The model predicts human perf.

- $d'$ ~ standardized error rate
- The higher the $d'$, the better the performance

Serre, Oliva, Poggio, 2005, PNAS, in press
The model matches human perf. (at this ISI)

Model 82% correct vs. humans 80%

High correlation of correct answers and errors between humans and computer model: \( \rho \approx 0.71 \)

Serre, Oliva, Poggio, 2005, PNAS, in press
Hierarchical models for the ventral stream: what is next?

- **A challenge for physiology:** disprove basic aspects of the architecture

- **Many “details” to be checked and filled out:**
  - Divisive normalization for tuning in V4 and IT
  - Unsupervised, developmental learning in V1, V4, PIT? How?
  - “Simple” cells in V4 or V2?
  - Extend to time-varying stimuli
  - Colour, stereo, grouping/segmentation...
Fun predictions to test: for instance... lesioning the model

Key in explaining the model results is:

- the large dictionary of features from V1 to PIT
- the loose hierarchy (i.e., different levels of selectivity and invariance)
Hierarchical models for the ventral stream: limitations

Recognition in clutter (see talk by Rosenholtz) is increasingly difficult
Performance of the model

- High performance (~90%) in the absence of clutter
- Performance decreases (~74%) with increasing amount of clutter
- Limitation of the feedforward model compatible with decrease in response in IT in the presence of clutter (Zoccolan, Cox, DiCarlo, 2005; Rolls, Aggelopoulos, Zheng, 2003)
Useful to remind the reading-out results in AIT: relevant for rapid categorization

77 isolated objects, 8 classes

Result relevant for rapid categorization discussion:

very rapid read-out of object information (irrespectively of identification or categorization), over very short times
We can decode from model units as well as from IT. For isolated objects model agrees with IT....
What about clutter? Recognition in “clutter” becomes increasingly difficult for feedforward architectures: *this needs to be tested with read-out in IT*...
Hierarchical models for the ventral stream: limitations

Recognition in clutter is increasingly difficult

→ Need for attentional bottleneck (Wolfe) perhaps in V4 (see Gallant and Desimone and models by Walther+Serre)?

→ Notice: this is a “novel” justification for the need of attention!
Bottleneck in V4 controlled by backprojections/attention: modeling to be done in close collaboration with physiologists
Beyond 50 ms: Feedback loops

- Prelim. results with model that includes spatial attention

Performance ($d'$) vs. SOA

Perf. of focused classifier close to human for SOA = 80ms

Serre, preliminary
Hierarchical models for the ventral stream: what is next?

- More needs to be added to feedforward architectures:
  - segmentation/grouping, colour, stereo, realistic synapses and spiking neurons, time (!)...
- Extension to include backprojections for $t > 50$ msec, for recognition in clutter and - more importantly - for image inference: recognition is more than scene categorization!!!
What is next?

More and more of the synergies we have seen in this SUnS workshop between psychophysics, computer vision and neuroscience!

😊