Deep learning and human vision

Mini lecture 2: hand-wired vs. learning

local building blocks

continuous valued inputs and outputs representing frequency of action potentials (spikes)

what determines the weights $w_i$?,
shallow convolutional networks
what can they do?

- detect edges
- detect faces
- detect textures

...but despite much research, few of these shallow networks work very well as models of human perception, except for simple stimuli and tasks
  e.g. they work well as predictors of contrast detection, discrimination of certain types of textures.

simple and complex cells in V1

one possible circuit model for complex cell—illustrating
local translation invariance
what determines the weights $w_{ij}$ as one proceeds up levels ($j$) of the hierarchy?,

the tasks of vision, e.g. “core” recognition

the regularities in images, e.g. high correlations between nearby pixels

hierarchical models for feature extraction given task constraints, e.g. core recognition

• Local features progressively grouped into more structured representations
• edges => contours => fragments => parts => objects
• Selectivity/invariance trade-off
• Increased selectivity for object/pattern type
• Decreased sensitivity to view-dependent variations of translation, scale and illumination

Fukushima 1988


simple and complex cells in V1

one model illustrating local translation invariance

simple & complex cells in V1

- Simple cells
  - “template matching”, i.e. detect conjunctions, logical “AND”

- Complex cells
  - insensitivity to small changes in position, detect disjunctions, logical “OR”

- Recognition as the hierarchical detection of “disjunctions of conjunctions”

Recognize the letter “†”

“†” is represented by the conjunction of a vertical and horizontal bar

\[
\text{AND} \quad \text{OR}
\]

which can occur at any one of many locations \(i\)

\[
\text{“†”}: \quad h_1 \& \& v_1 \ || \quad h_2 \& \& v_2 \ || \quad h_3 \& \& v_3... 
\]

volunteers to lead next week paper discussions?


