CS375 / Psych 249: Large-Scale Neural Network Models for Neuroscience

Lecture 3: Deep CNNs and the Ventral Visual stream
2018.10.01

Daniel Yamins
Stanford Neurosciences Institute
Stanford Artificial Intelligence Laboratory
Departments of Psychology and Computer Science
Stanford University
Problem: Entity Extraction

All the data I will show you today

Typical primate fixation duration during natural viewing

Chance is 50%
Adapted from DiCarlo et al. 2012
**GOAL:** Predictive model of single-neuron responses throughout the ventral stream to arbitrary image stimuli.
Hierarchical Convolutional Neural Networks

- Linear-Nonlinear neurally-plausible basic operations within layer

Layer components are basic neural-like operations.

**neuro:**
- synaptic weights patterns
- untangling through dimension expansion

**data:**
- single-unit activations
- complex cells
- competitive inhibition
- "AND" operation by limiting dynamic range
- adding robustness by dimension reduction
- put results back into standard range

Stimulus \( \rightarrow \) Neurons \( \rightarrow \) Behavior

Stimulus representation read-out

Layer components are basic neural-like operations.

Linear-Nonlinear neurally-plausible basic operations within layer

Madame Curie!
Neural Fitting Strategy?

Huge number of parameters consistent with HCNN concept.

i. **architectural** params: (# layers, # filters, receptive field sizes, &c) — “network structure”

ii. **filter** parameters: continuous valued pattern templates — “network contents”

*Q: How to discover the “right” parameters to understand real cortex?*
Models of V1

Where did this come from?

→ can just “see” what the right axes for measuring good tuning curves are, if we’re smart enough

(1) “Hubel and Wiesel’s Intuition”
~1970s and formalized later

Gaussian tuning curve of V1 simple cell

adapted from Adrienne Fairhall
Sparse Coding Foldiak, Olshausen, mid 1990s

$\mathbf{L}(\mathbf{x}) = |\mathbf{x} - O(H(\mathbf{x}))|^2 + \lambda \cdot |H(\mathbf{x})|$

→ neurons have to represent their environment, as efficiently as possible
Bottom-Up “Tuning Curve” Approach

V2:

- Original photographs
- V1-like filters matched: spectrally matched noise
- Correlations matched: naturalistic texture

So, maybe a hierarchically-built sparse auto-encoding in a 2-layer model with max pooling?? … but doesn’t really work well in practice.

V4:

Problem: No predictions for any other images.
  i.e.
  is not an “image-computable” model
Bottom-Up “Tuning Curve” Approach

IT:
Bottom-Up “Tuning Curve” Approach

(1) “Hubel and Wiesel’s Intuition”
   ~1970s and formalized later
   → e.g. there is a “fixed basis set”
   that just “makes sense” if we’re smart enough
   REALLY HARD TO GENERALIZE
   TO MULTI-LAYER NETWORKS

(2) Sparse Coding Foldiak, Olshausen,
    mid 1990s
   → neurons have to represent their environment, as efficiently as possible
Obvious alternative strategy: fit parameters to neural data.
Neural Fitting Strategy?

Huge number of parameters consistent with HCNN concept.

...not enough neural data to constrain model class. Gallant (2007); Rust & Movshon (2006)
Neural Fitting Strategy?

Huge number of parameters consistent with HCNN concept.

...not enough neural data to constrain model class. Gallant (2007); Rust & Movshon (2006)

Overfitting.
Optimize for Performance, Test Against Neurons

Visual Recognition Task

Step 1: Optimize for Task

Layer 1
Layer 2
Layer 3
Layer 4

Spatial Convolution over Image Input
Optimize for Performance, Test Against Neurons

Step 1: Optimize for Task

Spatial Convolution over Image Input

Step 2: Compare to Neural Data

Visual Recognition Task
1. **Performance**: accuracy on a challenging, high-variation visual object categorization task.

2. **Neural predictivity**: the ability of model to predict each individual neural site’s activity.
1. **Performance**: accuracy on a challenging, high-variation visual object categorization task.

2. **Neural predictivity**: the ability of model to predict each individual neural site's activity.

*challenging for neural network engineers, not the animal*
1. **Performance**: accuracy on a challenging, high-variation* visual object categorization task.

2. **Neural predictivity**: the ability of model to predict each individual neural site's activity.

Our hypothesis: Performance (1) → neural predictivity (2).

*challenging for neural network engineers, not the animal*
Multi-array electrophysiology in macaque V4 and IT.
(somewhere between single and multi-unit recording)
Multi-array Electrophysiology Experiment

5760 images

64 objects

8 categories

Uncorrelated photo backgrounds

Animals
Boats
Cars
Chairs
Faces
Fruits
Planes
Tables

Low variation
Medium variation
High variation

Pose, position, scale, and background variation
Multi-array Electrophysiology Experiment

About 300 total sites

Output = Binned spike counts 70ms-170ms post stimulus presentation averaged over 25-50 reps of each image.
Multi-array Electrophysiology Experiment

Responses to 1,600 test images of two example units

IT unit 53

Images sorted first by **category**, then **variation level**.
IT Neurons Track Human Performance

V4 loses out at higher variation:

... but humans are much less affected.

... as is the IT neural population.

At high variation levels, IT much better than V4 and existing models.

Yamins* and Hong* et. al. *PNAS* (2014)
Neural predictivity: the ability of model to predict each individual neural site’s activity.
Neural Response Prediction

Some kind of mapping is necessary.
Here, we use linear regression.

**Neural Response Prediction**

$$T = M \times S$$
Neural predictivity: the ability of model to predict each individual neural site’s activity.

Neural site unit ~ sparse linear combination of model units

Linear regression with fixed training images.

Accuracy = goodness-of-fit on held-out testing images (Cross validated)

Neural predictivity = median accuracy over all units.
High-throughput experiments to directly test the relationship between performance and IT neural predictivity.

- Random selection of model parameters; measure performance and neural predictivity  

  Pinto et. al (2008, 2009)
Initial Validation of Idea

random $r = 0.55 \pm 0.08$  
$n=2000$

Yamins* and Hong* et. al. PNAS (2014)
High-throughput experiments to directly test the relationship between neural predictivity and performance.


Initial Validation of Idea

performance-optimized

$r = 0.79 \pm 0.05$

$(n=2000)$
Initial Validation of Idea

High-throughput experiments to directly test the relationship between neural predictivity and performance.

- Random selection of model parameters; measure performance and neural predictivity  
  Pinto et al. (2008, 2009)

- Optimize parameters for performance; measure neural predictivity  
  optimization techniques: Bergstra Yamins & Cox (2013)

- Optimize parameters for neural predictivity; measure performance
Performance vs IT predictivity: Predictivity-Optimized

predictivity-optimized
\[ r = 0.80 \pm 0.04 \]
\( n=2000 \)

Yamins* and Hong* et. al. *PNAS* (2014)
Performance vs IT predictivity: Predictivity-Optimized

Performance is a potentially very good driver of neural prediction.

\[ r = 0.55 \pm 0.08 \]
\[ r = 0.79 \pm 0.05 \]
\[ r = 0.80 \pm 0.04 \]

Yamins* and Hong* et. al. PNAS (2014)
Performance vs IT predictivity

Yamins* and Hong* et. al. PNAS (2014)
Performance vs IT predictivity

But, not doing that well. Really want to be here:
Optimization Strategy

i. **architectural** params: (# layers, # filters, receptive field sizes, &c) — “network structure”

→ Automated meta-parameter optimization in high-dimensional discrete parameter spaces
   Bergstra Yamins & Cox (2013)

→ Ensembles of models chosen through modified boosting  Yamins et. al (2013, 2014)
Optimization Strategy

i. Architectural params: (# layers, # filters, receptive field sizes, &c) — “network structure”

→ Automated meta-parameter optimization in high-dimensional discrete parameter spaces

Bergstra Yamins & Cox (2013)

→ Ensembles of models chosen through modified boosting

Yamins et. al (2013, 2014)

ii. Filter parameters: continuous valued pattern templates — “network contents”

→ GPU-accelerated stochastic gradient descent

Pinto et. al., (2009), Krizhevsky et. al. (2012)

Gradient descent eq: \[
\frac{dp}{dt} = -\lambda(t) \cdot \frac{\partial L}{\partial P}
\]

L = loss function
\lambda = learning rate

In current practice:

L = loss computed from large numbers of externally-provided object category labels.
Model Training Regimen

**train:** real photos
Model Training Regimen

**train:** real photos

**test:** neural stimuli

generalize?
Model Training Regimen

**train:** real photos

**test:** neural stimuli

removed categories of photos that appeared in the test stimuli

(animals, boats, cars, chairs, faces, fruits, planes, tables)

**generalize?**
Basic categorization removed categories of photos that appeared in the test stimuli (animals, boats, cars, chairs, faces, fruits, planes, tables).

\[\text{train: real photos} \rightarrow \text{removed categories of photos that appeared in the test stimuli} \rightarrow \text{test: neural stimuli} \rightarrow \text{Specific 4-layer model that achieved high recognition performance.}\]
Performance Generalization

- Basic categorization
- Test performance vs. training time graph
At high variation levels, IT much better than V4 and existing models.

Performance Comparison

Yamins* and Hong* et. al. *PNAS* (2014)
At high variation levels, IT much better than V4 and existing models.

New model comparable to IT / human performance levels.
Performance Comparison

Yamins* and Hong* et. al. PNAS (2014)

Low Var. | Medium Var. | High Var.
--- | --- | ---
Pixels | SIFT | V1-like | V2-like | HMAX | PLOS09 | V4 Neurons | IT Neurons | Human | HMO |

Basic categorization
Car identification
Face identification
Behavioral match between models and data at category confusion level is pretty good …

Performance Comparison

![Graphs comparing human and model performance](image_url)

- **IT**
- **V4**
- **Model**
Does it predict neurons better?
Does it predict neurons better?

Images sorted first by **category**, then **variation level**.

Neural data

**Yamins* and Hong* et. al. PNAS (2014)**
Does it predict neurons better?

Images sorted first by category, then variation level.

Yamins* and Hong* et. al. *PNAS* (2014)

unit 53

- Neural data
- Model prediction
Predicting IT Neural Responses

Images sorted first by category, then variation level.

- Neural data
- Model prediction
Key Underlying Principle

IT Explained Variance (%)

$\text{Categorization Performance (balanced accuracy)}$

V2-like

HMAX

PLOS09

$\text{SIFT}$

$r = 0.87 \pm 0.15$

0.6

0.8

1.0

HMO

50

40

30

20

10

0

IT Explained Variance (%)

0

10

20

30

40

50

\(\text{V1-like} \quad \text{HMAX} \quad \text{PLOS09} \quad \text{SIFT} \quad \text{Pixels} \quad \bigcirc = \text{distinct model} \quad \text{HMO} \)

Yamins* and Hong* et. al. *PNAS* (2014)
Predicting IT Neural Responses

What about intermediate layers?

i. compare intermediate model layers to IT neural data

ii. compare all model layers to intermediate visual areas (V4)
Captures low variation image response patterns …
Layer 1

Captures low variation image response patterns …

Animals  Boats  Cars  Chairs  Faces  Fruits  Planes  Tables

- Neural data
- Model prediction
Layer 1

Animals  Boats  Cars  Chairs  Faces  Fruits  Planes  Tables

Neural data  Model prediction

... but fails to capture higher variation response patterns.
Layer 1

Neural data

Model prediction

Animals  Boats  Cars  Chairs  Faces  Fruits  Planes  Tables
Layer 3

Layer 2

Layer 1

Animals  Boats  Cars  Chairs  Faces  Fruits  Planes  Tables

Neural data  Model prediction
Building tolerance while maintaining selectivity

Layer 1

Layer 2

Layer 3

Top Layer

Animals  Boats  Cars  Chairs  Faces  Fruits  Planes  Tables
Predicting IT Neural Responses

Yamins* and Hong* et. al. *PNAS* (2014)
Predicting IT Neural Responses

Yamins* and Hong* et. al. PNAS (2014)
Predicting IT Neural Responses

---

Yamins* and Hong* et. al. *PNAS* (2014)
Predicting IT Neural Responses

Yamins* and Hong* et. al. PNAS (2014)
Predicting IT Neural Responses

Performance constraints + architectural constraints → better neural prediction
What about intermediate layers?

i. compare intermediate model layers to IT neural data

ii. compare all model layers to intermediate visual areas (V4)
Predicting V4 Neural Responses

V4 unit 60

Animals | Boats | Cars | Chairs | Faces | Fruits | Planes | Tables
Predicting V4 Neural Responses

V4 unit 60

Animals  Boats  Cars  Chairs  Faces  Fruits  Planes  Tables

Layer 1

Top Layer

Neural data  Model prediction
Predicting V4 Neural Responses

- Layer 1
- Layer 2
- Layer 3
- Top Layer

Animals | Boats | Cars | Chairs | Faces | Fruits | Planes | Tables

Neural data | Model prediction
Predicting V4 Neural Responses

Yamins* and Hong* et. al. *PNAS* (2014)
Investigating fits as a function of model layer:

IT fit increases at each layer. In contrast, V4 fit peaks and then goes down.

Yamins* and Hong* et. al. PNAS (2014)
Nothing special about 4 layers — deeper models can be better:

Hong* and Yamins* et al.

*Nature Neuroscience*

(2016)
Layer-area correspondence

Top hidden layer (not explicit categorization layer)

Hong* and Yamins* et. al.
*Nature Neuroscience (2016)
Layer-area correspondence

Top **hidden** layer (*not* explicit categorization layer)

Macaque ephys

Hong* and Yamins* et. al.  
*Nature Neuroscience*  
(2016)

Human fMRI

Khaligh-Razavi & Kriegeskorte (2014)
Layer-area correspondence

Top **hidden** layer (not explicit categorization layer)

- Macaque ephys

- *Hong* and *Yamins* et al. *Nature Neuroscience* (2016)

- Khaligh-Razavi & Kriegesmorde (2014)

**Best recent models:** \(~13\) layers deep, with IT best predicted around \(~80\%\) of the way through (e.g. 10 layers)
Layer-area correspondence

Model filters at lowest layer resembles Gabor wavelets:
Model filters at lowest layer resembles Gabor wavelets:

Model early layers are best explanation of fMRI data in V1. (with Darren Seibert and Justin Gardner)

Deep convolutional models improve predictions of macaque V1 responses to natural images

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, Alexander S Ecker

doi: https://doi.org/ 10.1101/201764
Deep convolutional models improve predictions of macaque V1 responses to natural images

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, Alexander S Ecker

doi: https://doi.org/10.1101/201764
Deep convolutional models improve predictions of macaque V1 responses to natural images

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolihas, Matthias Bethge, Alexander S Ecker

doi: https://doi.org/10.1101/201764

Layer-area correspondence

finer resolution
Layer-area correspondence

Deep convolutional models improve predictions of macaque V1 responses to natural images

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, Alexander S Ecker

doi: https://doi.org/10.1101/201764

50% explained variance vs
- 17% for Linear-Nonlinear-Poisson (with gabor filters)
- 39% for Berkeley Wavelet Transform

finer resolution
Layer-area correspondence

Deep convolutional models improve predictions of macaque V1 responses to natural images

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, Alexander S Ecker
doi: https://doi.org/10.1101/201764

50% explained variance vs
- 17% for Linear-Nonlinear-Poisson (with gabor filters)
- 39% for Berkeley Wavelet Transform

Peak V1 Peak V4 Peak IT
(unpublished) (unpublished)
Layer-area correspondence

Deep convolutional models improve predictions of macaque V1 responses to natural images

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, Alexander S Ecker

doi: https://doi.org/10.1101/201764

- 50% explained variance vs
  - 17% for Linear-Nonlinear-Poisson (with gabor filters)
  - 39% for Berkeley Wavelet Transform

Peak V1
Peak V4 (unpublished)
Peak IT (unpublished)
Peak V2??

subcortical??
Deep Learning Models of the Retinal Response to Natural Scenes

Lane T. McIntosh*1, Niru Maheswaranathan*1, Aran Nayebi1, Surya Ganguli2,3, Stephen A. Baccus3
1Neurosciences PhD Program, 2Department of Applied Physics, 3Neurobiology Department
Stanford University
{lmcintosh, nirum, anayebi, sganguli, baccus}@stanford.edu
Three-layer CNN best fits retinal ganglion cell response patterns to natural images.
Layer-area correspondence

State-of-the-art model of the ventral visual stream:

- V4 at 6th convolutional layer
- pIT at 7th convolutional layer
- cIT/aIT at layers 8-10, depending on neurons position on A/P axis
Behavioral “Top-Down” constraints

Complement standard “from below” approach …
Behavioral “Top-Down” constraints

Complement standard “from below” approach ...
Behavioral “Top-Down” constraints

Complement standard “from below” approach …
Behavioral “Top-Down” constraints

Complement standard “from below” approach …
Behavioral “Top-Down” constraints

Complement standard “from below” approach … with behavioral constraints
Behavioral “Top-Down” constraints

Complement standard “from below” approach … with behavioral constraints
Behavioral “Top-Down” constraints

Complement standard “from below” approach … with behavioral constraints
Behavioral “Top-Down” constraints

Complement standard “from below” approach … with behavioral constraints
Beyond categorization

**Category**

**Identity**

plane

f16
Beyond categorization

Position
Beyond categorization

Size
Beyond categorization

Aspect Ratio and Angle
Beyond categorization

We can quickly assess the scene as a whole.
Where and how are all these properties coded neurally?
Beyond categorization

“Standard word model” predicts: **not at the top of the ventral stream.**

Aggregation over identity-preserving transformations, e.g. translation.
Beyond categorization

“Standard word model” predicts: **not at the top of the ventral stream.**

Aggregation over identity-preserving transformations, e.g. translation.

![Diagram showing receptive field size and category invariance increasing with higher brain areas](image)
Beyond categorization

“Standard word model” predicts: not at the top of the ventral stream.

Aggregation over identity-preserving transformations, e.g. translation.

- Receptive Field Size $\uparrow$
- Category Invariance $\uparrow$
- (e.g.) Position Sensitivity $\downarrow$

position / size estimation

categorization

IT

pose?

V4

V2

V1
Where and how are all these properties coded neurally?

dorsal stream?

earlier visual areas?

V1
V2
V4
IT
Somewhat newish ideas about IT?

State of knowledge from previous studies . . .

Depth Along Ventral Stream

*(increasing receptive field size →)*
Somewhat newish ideas about IT?

State of knowledge from previous studies . . .

Population Decode Performance
(relative to human performance)

Depth Along Ventral Stream
(increasing receptive field size →)

Multiple hypotheses consistent with the existing data . . .

H1: Tolerance / sensitivity tradeoff?
Somewhat newish ideas about IT?

State of knowledge from previous studies . . .

Categorization

Orthogonal Properties

Multiple hypotheses consistent with the existing data . . .

H1

H2

H3

H3: Information preservation?

Population Decode Performance (relative to human performance)

Depth Along Ventral Stream (increasing receptive field size →)

Population Decode Performance
Unexpected observation:

Training on categorization task Increased performance on position estimation task.

even though the goal was to become INVARIANT to position

Beyond categorization

Category optimization → improved performance on non-categorical tasks.

Beyond categorization

Unexpected observation #2:

Test Performance

Categorization

Model Layers

Layer 1 Layer 2 Layer 3 Layer 4 Layer 5 Layer 6

Unexpected observation #2:

*Increased* performance on position estimation task at each model layer.

Beyond categorization

For all tasks of visual interest we could measure in our test dataset:

Performance on non-categorical tasks increases at each layer:
Beyond categorization

What do the data say?
Population Decoding

Population Decoding

**IT > V4, V1** for all tasks  **V4 > V1** for most tasks

- **Categorization**
- **Identification**

<table>
<thead>
<tr>
<th>Feature</th>
<th>IT cortex</th>
<th>V4 cortex</th>
<th>V1-like model</th>
<th>Pixel control</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-axis Position</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y-axis Position</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X-axis Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y-axis Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bounding Box Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-D Retinal Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perimeter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-D Object Scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Major Axis Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspect Ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Major Axis Angle</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z-axis Rotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y-axis Rotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X-axis Rotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Single Site Responses

Best single position-encoding sites.

heat map value at \( x, y = \)
response averaged over all images where object center is in position \( x, y \)
Single Site Responses

Best single position-encoding sites.

heat map value at \( x, y = \) response averaged over all images where object center is in position \( x, y \)

Population Decoding

“Standard” receptive field-mapping stimuli w/ position and orientation variation:
Population Decoding

**V1 > V4, IT** for “standard” tasks

*IT cortex*, *V1-like model*, *V4 cortex*, *pixel control*

Human Psychophysical Measurements

Click where the boat was!

6 learning trial(s) left after this.
Monkey Neurons vs Humans

performance \sim k \times \log(N)

Basic Categorization

Subordinate Identification

fraction of human performance

number of neural sites

Monkey Neurons vs Humans

## Monkey Neurons vs Humans

<table>
<thead>
<tr>
<th></th>
<th>IT</th>
<th>V4</th>
<th>V1</th>
<th>Pix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Categorization</td>
<td>773 ± 185</td>
<td>2.2 x 10^6</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Subordinate Identification</td>
<td>496 ± 93</td>
<td>4.4 x 10^6</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>X-axis Position</td>
<td>1414 ± 403</td>
<td>5.2 x 10^5</td>
<td>3.0 x 10^7</td>
<td>—</td>
</tr>
<tr>
<td>Y-axis Position</td>
<td>918 ± 309</td>
<td>2.5 x 10^4</td>
<td>8.7 x 10^6</td>
<td>—</td>
</tr>
<tr>
<td>Bounding Box Size</td>
<td>322 ± 90</td>
<td>1.7 x 10^4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>X-axis Size</td>
<td>256 ± 87</td>
<td>9.8 x 10^3</td>
<td>3.4 x 10^7</td>
<td>—</td>
</tr>
<tr>
<td>Y-axis Size</td>
<td>237 ± 87</td>
<td>3.8 x 10^3</td>
<td>9.5 x 10^6</td>
<td>—</td>
</tr>
<tr>
<td>3-D Object Scale</td>
<td>401 ± 90</td>
<td>3.2 x 10^4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Major Axis Length</td>
<td>201 ± 70</td>
<td>1.1 x 10^4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Aspect Ratio</td>
<td>163 ± 61</td>
<td>951 ± 59</td>
<td>6.5 x 10^3</td>
<td>—</td>
</tr>
<tr>
<td>Major Axis Angle</td>
<td>804 ± 136</td>
<td>3.2 x 10^6</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Z-axis Rotation</td>
<td>1932 ± 1061</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Y-axis Rotation</td>
<td>369 ± 115</td>
<td>2.8 x 10^5</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>X-axis Rotation</td>
<td>1570 ± 530</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

— = more than 10 billion sites required


Mean over tasks, human-parity for IT is at ~700 multi-unit trial-averaged sites.
## Monkey Neurons vs Humans

<table>
<thead>
<tr>
<th></th>
<th>IT</th>
<th>V4</th>
<th>V1</th>
<th>Pix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Categorization</td>
<td>773 ± 185</td>
<td>2.2 × 10^6</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Subordinate Identification</td>
<td>496 ± 93</td>
<td>4.4 × 10^6</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>X-axis Position</td>
<td>1414 ± 403</td>
<td>5.2 × 10^5</td>
<td>3.0 × 10^7</td>
<td>—</td>
</tr>
<tr>
<td>Y-axis Position</td>
<td>918 ± 309</td>
<td>2.5 × 10^4</td>
<td>8.7 × 10^6</td>
<td>—</td>
</tr>
<tr>
<td>Bounding Box Size</td>
<td>322 ± 90</td>
<td>1.7 × 10^4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>X-axis Size</td>
<td>256 ± 87</td>
<td>9.8 × 10^3</td>
<td>3.4 × 10^7</td>
<td>—</td>
</tr>
<tr>
<td>Y-axis Size</td>
<td>237 ± 87</td>
<td>3.8 × 10^3</td>
<td>9.5 × 10^6</td>
<td>—</td>
</tr>
<tr>
<td>3-D Object Scale</td>
<td>401 ± 90</td>
<td>3.2 × 10^4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Major Axis Length</td>
<td>201 ± 70</td>
<td>1.1 × 10^4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Aspect Ratio</td>
<td>163 ± 61</td>
<td>951 ± 59</td>
<td>6.5 × 10^3</td>
<td>—</td>
</tr>
<tr>
<td>Major Axis Angle</td>
<td>804 ± 136</td>
<td>3.2 × 10^6</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Z-axis Rotation</td>
<td>1932 ± 1061</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Y-axis Rotation</td>
<td>369 ± 115</td>
<td>2.8 × 10^5</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>X-axis Rotation</td>
<td>1570 ± 530</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

— = more than 10 billion sites required

Mean over tasks, human-parity for IT is at ~350000 single-unit single-trial neurons.

Somewhat newish ideas about IT?

State of knowledge from previous studies . . .

Multiple hypotheses consistent with the existing data . . .

Depth Along Ventral Stream
(increasing receptive field size →)

H4: Simultaneous build-up of encoding
Somewhat newish ideas about IT?

1. IT is *NOT* invariant. Strict generalization of simple-to-complex cells: **no.**

2. “Lower-level” properties are not that low-level — at least, with complex objects and backgrounds.

3. Categorization and non-categorical properties “go together” — *not* just that “not all (e.g.) position information is lost” (MacEvoy 2013, DiCarlo 2003)

Provides support to a hypothesis for what IT does:

“Inverting the generative model of the scene”
Principle of “Goal-Driven Modeling”
Principle of “Goal-Driven Modeling”
Heuristic of “Goal-Driven Modeling”
“Mercedes behind Lamborghini, on a field in front of mountains.”

“Hannah is good at compromising”
Heuristic of “Goal-Driven Modeling”

“Mercedes behind Lamborghini, on a field in front of mountains.”

“Hannah is good at compromising”
But what type of understanding is this?
But what type of understanding is this?

_not saying this type of understanding is impossible…_
> Formulate comprehensive model class (CNNs)

Choose challenging, ethologically-valid tasks (categorization)

Formulate comprehensive model class (CNNs)

> Formulate comprehensive model class (CNNs)

> Choose challenging, ethologically-valid tasks (categorization)

> Implement generic learning rules (gradient descent)

> Formulate comprehensive model class (CNNs)

> Choose challenging, ethologically-valid tasks (categorization)

> Implement generic learning rules (gradient descent)

> Map to brain data. (ventral stream)
A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons

David Zipser* & Richard A. Andersen†

* Institute for Cognitive Science, University of California, San Diego, La Jolla, California 92093, USA
† Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Neurons in area 7a of the posterior parietal cortex of monkeys respond to both the retinal location of a visual stimulus and the position of the eyes and by combining these signals represent the spatial location of external objects. A neural network model, programmed using back-propagation learning, can decode this spatial information from area 7a neurons and accounts for their observed response properties.

(Nature, 1988)
Optimize for Performance, Test Against Neurons

Area 7a in context
3-layer network (one hidden layer) with weights optimized to solve task of getting head-coordinate outputs from retinal input and eye coordinate

Zipser and Andersen 1988
3-layer network (one hidden layer) with weights optimized to solve task of getting head-coordinate outputs from retinal input and eye coordinate

inputs modeled on neurons in 7a with “no eye-position dependence”

Zipser and Andersen 1988
Optimize for Performance, Test Against Neurons

outputs perhaps describe 7a neurons *with* eye-position dependence

3-layer network (one hidden layer) with weights optimized to solve task of getting head-coordinate outputs from retinal input and eye coordinate

inputs modeled on neurons in 7a with “no eye-position dependence”

Zipser and Andersen 1988
Optimize for Performance, Test Against Neurons

Fig. 6  Hidden unit spatial gain fields generated by the model network. Fields a–f were generated using the monotonic format output; the rest used the gaussian format output.

Zipser and Andersen 1988
Optimize for Performance, Test Against Neurons

Fig. 2 The receptive fields of spatially tuned neurons from area 7a, arranged in rows with the eccentricity of the field maxima increasing to the right, and in columns with the complexity of the fields increasing downwards. Receptive fields were sampled at 17 radially spaced points, with one sample taken at the centre of the field, and four samples taken on each of four circles of radius 10, 20, 30 and 40 degrees. All the fields in row a have single peaks. Those in row b have a single large peak but some complexities in the field. The fields in row c are the most complex with multiple peaks. The data have been normalized so that the highest peak in each field is the same height.
Fig. 5  Hidden unit retinal receptive fields generated by the back propagation model. These plots were generated by holding the eye-position.
Fig. 2 The receptive fields of spatially tuned neurons from area 7a, arranged in rows with the eccentricity of the field maxima increasing to the right, and in columns with the complexity of the fields increasing downwards. Receptive fields were sampled at 17 radially spaced points, with one sample taken at the centre of the field, and four samples taken on each of four circles of radius 10, 20, 30, and 40 degrees. All the fields in row a have single peaks. Those in row b have a single large peak but some complexities in the field. The fields in row c are the most complex with multiple peaks. The data have been normalized so that the highest peak in each field is the same height.

Fig. 5 Hidden unit retinal receptive fields generated by the back propagation model. These plots were generated by holding the eye position for the experimental receptive fields (Fig. 2). Comparison of the top lines in Figs 2 and 5 shows that the trained models generate single-peak receptive fields resembling those observed experimentally at all eccentricities except 10 degrees. The fully trained model also produces moderately complex fields like those found in line 2, but rarely produces receptive fields as complex as those in the bottom line of Fig. 2. This kind of highly complex field is not distinguishable from the untrained model receptive fields shown in the bottom of Fig. 5. The comparison process contains an element of subjectivity, but it demonstrates that the trained model generates retinal receptive fields remarkably similar to the experimentally observed fields. The gain fields
\[
\arg\min_{a \in A} [L(p_a^*)]
\]

where \( p^* \) is result of

\[
\frac{dp_a}{dt} = -\lambda(t) \cdot \langle \nabla p_a L(x) \rangle_{x \in D}
\]

\( A = \text{architecture class} \)
\( L = \text{loss function} \)
\( D = \text{dataset} \)
1. \[ A = \text{architecture class} \]

2. \[ L = \text{loss function} \quad D = \text{dataset} \]

3. \[ \argmin_{a \in A} [L(p_a^*)] \]

where \( p^* \) is result of

\[ \frac{dp_a}{dt} = -\lambda(t) \cdot \langle \nabla p_a L(x) \rangle_{x \in D} \]

“learning rule”

“task”
Principle of “Goal-Driven Modeling”

Heuristic of “Goal-Driven Modeling”

… after all at some point, for any given task, you’ll probably “go over the hump” … perhaps when you exceed human performance or overfit on that task.
> Formulate comprehensive model class (**CNNs**)

> Choose challenging, ethologically-valid tasks (**categorization**)

> Implement generic learning rules (**gradient descent**)

> Map to brain data. (**ventral stream**)

> Formulate comprehensive model class (**RNNs**)

> Choose challenging, ethologically-valid tasks (**task switching/memory**)

> Implement generic learning rules (**??**)

> Map to brain data. (**Parietal cortex, PFC**)