Brain Learning and Recognition: The Large and the Small of It in Inferotemporal Cortex

Grossberg, Stephen

Boston University Center for Adaptive Systems and Department of Cognitive and Neural Systems

http://hdl.handle.net/2144/1973

Boston University
Brain Learning and Recognition:
The Large and the Small of It in Inferotemporal Cortex

Stephen Grossberg, Jeffrey Markowitz, and Yongqiang Cao

Department of Cognitive and Neural Systems
Center for Adaptive Systems
Center of Excellence for Learning in Education, Science and Technology
Boston University
677 Beacon Street, Boston, MA, 02215, USA
Email:{steve, jmarkow, yqcao}@cns.bu.edu

All correspondence should be addressed to
Professor Stephen Grossberg
Department of Cognitive and Neural Systems
Boston University
677 Beacon Street
Boston, MA 02215
Phone: 617-353-7858/7
Fax: 617-353-7755

Submitted: June 22, 2009


Acknowledgements: Supported in part by CELEST, an NSF Science of Learning Center (SBE-0354378), and by the SyNAPSE program of the Defense Advanced Research Projects Agency (HR0011-09-3-0001 and HR0011-09-C-0011).

Copyright © 2009

Permission to copy without fee all or part of this material is granted provided that: 1. The copies are not made or distributed for direct commercial advantage; 2. the report title, author, document number, and release date appear, and notice is given that copying is by permission of the BOSTON UNIVERSITY CENTER FOR ADAPTIVE SYSTEMS AND DEPARTMENT OF COGNITIVE AND NEURAL SYSTEMS. To copy otherwise, or to republish, requires a fee and / or special permission.
Abstract

Anterior inferotemporal cortex (ITa) plays a key role in visual object recognition. Recognition is tolerant to object position, size, and view changes, yet recent neurophysiological data show ITa cells with high object selectivity often have low position tolerance, and vice versa. A neural model learns to simulate both this tradeoff and ITa responses to image morphs using large-scale and small-scale IT cells whose population properties may support invariant recognition.
Classical lesion studies have shown that inferotemporal cortex (IT) supports visual object recognition in primates\(^1\). Single cell recordings show that IT cells have large receptive fields (~23°), that contain the fovea and ipsi- and contra-lateral visual fields, and respond selectively to ‘complex’ objects; e.g. two-dimensional silhouettes of hands and Fourier descriptors\(^2,3\). These properties are natural in a cortical area subserving invariant object recognition. However, recent experiments show that some IT cells have much smaller receptive fields\(^4\), and that ITa cells exhibit a *tradeoff* between object selectivity (or sparseness) and position tolerance (or invariance) wherein neurons with high object selectivity typically have low position tolerance and vice versa\(^5\).

Figure 1: LEFT, Model circuit: LGN ON (OFF) cells map to cortex via the log-polar transform and are then boundary processed with simple (half filled) and complex (not filled) cells. Then, an attentive ART category learning and recognition network models responses of ITa cells. RIGHT, ART circuit: Top-down expectations are attentively matched against bottom-up input features. Attended critical features are learned. A big enough mismatch (i.e., one that does not satisfy vigilance in the orienting system O) causes a burst of novelty-sensitive arousal from O that inhibits the active category and triggers search for and learning of a better matching category. C: biased competition; open triangles: inhibitory connections; filled triangles: excitatory connections; filled semi-circles: learned connections.

We propose a neural model (Figure 1; see Supplementary Materials) that quantitatively explains and simulates these data, among others. Some prominent efforts to model IT have built invariant representations using a hierarchy of feedforward filters leading to a learned category choice\(^6-8\), or
through grouping object translations through time\textsuperscript{6,9,10}. Our model goes beyond them by clarifying how multiple-scale processing, when combined with experimentally documented ON and OFF cell pre-processing, contrast normalization, cortical magnification factor, boundary processing, and attentionally-modulated object category learning, can naturally explain the observed tradeoff and is compatible with invariant recognition.

The model learned from natural objects, chosen from the Cal Tech 101 image database\textsuperscript{11} that was used by Zoccolan et al. (2007). Each image was processed by ON and OFF cells with three receptive field sizes, from coarse to fine. The model LGN contrast-normalized the images using on-center off-surround networks whose cells obey membrane, or shunting, equations\textsuperscript{12}. These contrast-normalized activity patterns at each spatial scale then underwent a log-polar transform that computes the cortical magnification factor in striate cortex, which over-represents the fovea and under-represents the periphery, a critical point since Zoccolan et al. (2007) presented images up to 10° from the fovea.

Then, the contrast-normalized and cortically magnified images in each scale generated boundary representations that were computed by oriented, multiple-scale, contrast-polarity-sensitive difference-of-offset-Gaussian filters, which combined ON and OFF cell signals via a self-normalizing shunting network that models cortical area V1 simple cells. Then opposite contrast-polarity-sensitive simple cell outputs were combined at contrast-invariant complex cells whose signals input to the model ITa.

Model ITa embodies incremental, fast, learning of recognition categories whose bottom-up filters and learned top-down expectations learn prototypes of attended critical features using biased competition\textsuperscript{12,13}. Learning in an Adaptive Resonance Theory, or ART, network is modulated by a vigilance parameter (\(\rho\)), which controls the generality of the learned categories: low vigilance leads to general, abstract categories (e.g., general face category); high vigilance leads to specific, concrete categories (e.g., view of a single face)\textsuperscript{12}. Neurobiological, cognitive, and clinical correlates of vigilance control are reviewed in \cite{14}. Since the monkeys in the Zoccolan et al. experiment passively experienced the stimuli, vigilance was set low (\(\rho\) equal to .1).

The 203 simulated Cal Tech 101 images included stimuli from the experiment at, below, and above the center of gaze. For each run, 200 random samples without replacement were used during learning, during which ART attentional and orienting mechanisms discovered and learned new categories. During testing, all 203 objects were presented at the center to measure selectivity, and 49 objects were presented at all three training positions to measure position tolerance.

From four simulations of 10 runs, we calculated the mean correlation coefficient and standard deviation [in brackets] between tolerance and selectivity: (1) -.41 [.119]; (2) -.39 [.123]; (3) -.45 [.105]; and (4) -.42 [.160]. These results closely fit the correlations in Zoccolan et al. (2007); namely, -.39 (see Figure 2a-b). Multiple-scale processing, followed by attentive, vigilance-modulated ART category learning and recognition, were key to quantitatively simulating the tradeoff between selectivity and tolerance. Moreover:

1) cells with a large spatial scale tend to exhibit low object selectivity (or sparseness) and high tolerance (or invariance), whereas

2) cells with a small spatial scale tend to exhibit high object selectivity but low tolerance.

The model also simulates responses of ITa cells to image morphs\textsuperscript{15} (see Figure 2c-d). Here monkeys did active discrimination during delayed matching-to-sample. We therefore set vigilance high (\(\rho\) equal to .9).
In summary, the model quantitatively explains key neurophysiological data concerning how ITa cells may generate the selectivity/tolerance tradeoff and graded responses to image morphs. Significantly, the selectivity/tolerance tradeoff was recorded during passive viewing conditions. During active scanning of objects with saccadic eye movements, spatial attentional mechanisms in the parietal cortex may modulate object category learning in ITa to bind multiple object views into a more position- and view-invariant object category representation. This suggests testing the selectivity/tolerance tradeoff both before and after active scanning and view-invariant object learning take place.
Supplementary Material

I. Stimulus Set

Figure S1: All 203 images were adapted from the Cal Tech 101 database, as were the stimuli used in the Zoccolan et al.\textsuperscript{5} experiment.
II. Simulation Procedure

We constructed a master dataset of images from the Cal Tech 101 database (see Figure S1) scaled to approximately 100 x 100 pixels and presented against a uniform white background of 300 x 300 pixels\textsuperscript{11}. Images were superimposed over the background at: (1) the center, (2) 50 pixels below the center, or (3) 50 pixels above the center. This resulted in a master dataset of 609 images (203 images x 3 positions). For the learning phase of each run, 200 images were randomly drawn from the master dataset without replacement and new model ITa recognition categories were incrementally learned by a variant of Adaptive Resonance Theory (ART), called Fuzzy ART, which is described below. During the testing phase, we froze learning while presenting two fixed subsets of the master dataset: (1) all 203 objects presented at the center to measure selectivity and (2) 49 objects presented at all three positions to measure position tolerance (49 objects x 3 positions = 147 total images). In both phases, each input image was presented to the network one at a time.

III. Model Overview

a) Preprocessing

Each input image is a circular ‘cut-out’ of the original image, which is then split into oversampled hemi-retinas (see Figure S2). Oversampling includes part of the opposing hemi-retina near the vertical meridian, consistent with neurophysiological evidence\textsuperscript{16}. This process is described in further detail in [10]. Cell processing begins by taking the average of the three RGB color values of an input image\textsuperscript{17}, split into hemi-retinas:

\[
I_{pq}^h = \frac{1}{3} (I_{pq}^{hr} + I_{pq}^{hg} + I_{pq}^{hb})
\]  

(1)

Here, the indices \( p \) and \( q \) correspond to the Cartesian \( x \) and \( y \) coordinates, respectively, of the input image and are used as dummy indices below; \( h \) indexes the hemi-retina \{left, right\}; and \( R,G,B \) the intensity values in each color channel.

b) Retina/LGN Processing

The property of retinal and LGN cells that is modeled is contrast normalization. This is achieved using, at each of three scales \( g = 1,2,3 \), an on-center off-surround network whose cells obey membrane, or shunting, equations\textsuperscript{18}. The narrow on-center is defined by a single pixel and the off-surround by a Gaussian kernel, centered on location \((i,j)\) in the image, whose breadth varies with the scale\textsuperscript{19,20}. Solved at equilibrium, the activity, or potential, \( x_{iq}^{hg} \), of the ON cell at position \((i,j)\), hemi-retina \( h \), and scale \( g \) is:
The existence of multiple scales is supported by psychophysical and neurophysiological data\textsuperscript{21-24}, and helps to explain the selectivity/tolerance tradeoff observed by Zoccolan et al.\textsuperscript{5}. The outputs of the ON and OFF channels were:

\[
X_{ij}^{bg+} = [X_{ij}^{bg+}]^+, \\
X_{ij}^{bg-} = [-X_{ij}^{bg+}]^+. 
\] (4)
where \([x]^+ = \max(x, 0)\) denotes a half-wave rectifying output signal, and + and – refer to the ON and OFF channels, respectively.

c) Log-polar Map

The ON and OFF channels of each hemi-retina undergo a log-polar transform that maps from retina to cortex, with the left hemi-retina forming the right hemispheric image and vice-versa for the left hemisphere:

\[
\begin{align*}
z^{hc} &= re^{i\theta} \\
w^{hc} &= \log(z^{hc} + a).
\end{align*}
\]

In (5), \(z\) is a complex number formed by the retinal image in polar coordinates, \(h\) refers to the hemi-retina, \(c\) the ON and OFF channels, and \(a\) is a constant (.7). This transform represents the cortical magnification factor in humans and other primates. The output of this operation is a log-polar map of each hemi-retina, \(V_{\text{left},c}\) and \(V_{\text{right},c}\). Both of them are then concatenated into a single image \(W^c\), where \(c\) designates either the ON channel (c+) or the OFF channel (c-):

\[
\begin{align*}
W^{c+} &= (V_{\text{left},c+}, V_{\text{right},c+}) \\
W^{c-} &= (V_{\text{left},c-}, V_{\text{right},c-})
\end{align*}
\]

d) Simple and Complex Cell Boundary Processing

Oriented multiple-scale simple cells are simulated in layer 4 of V1. The simple cells act as contrast-polarity-sensitive filters that detect oriented features in the image along the filter’s preferred orientation. In model simulations, cells selective for 4 different orientations are used within each of the 3 spatial scales.

V1 layer 4 simple cells are bottom-up activated by LGN ON and OFF outputs that are filtered by spatially elongated and offset Gaussian kernels. In particular, a layer 4 simple cell activity \(y^{g}_{ijk}\) at position \((i,j)\), orientation \(k\), and scale \(g\) obeys the shunting on-center off-surround equation:

\[
\frac{d}{dt} y^{g}_{ijk} = -\alpha y^{g}_{ijk} + (1 - y^{g}_{ijk}) \sum_{p,q} \left( W_{pq}^{c+} G_{pqijk}^{g+} + W_{pq}^{c-} G_{pqijk}^{g-} \right) \\
- (1 + y^{g}_{ijk}) \sum_{p,q} \left( W_{pq}^{c+} G_{pqijk}^{g-} + W_{pq}^{c-} G_{pqijk}^{g+} \right)
\]

In (8), the passive decay rate \(\alpha = 1\). In the excitatory term of (8), the log-polar transformed LGN ON cell output signals \(W_{pq}^{c+}\) are filtered by the oriented, spatially-elongated Gaussian kernel \(G_{pqijk}^{g+}\), while the LGN OFF output signals \(W_{pq}^{c-}\) are filtered by a similar kernel \(G_{pqijk}^{g-}\). The centers of the kernels \(G_{pqijk}^{g+}\) and \(G_{pqijk}^{g-}\) are offset in mutually opposite directions from each simple cell’s centroid along an axis perpendicular to the simple cell’s direction of elongated sampling. In the
inhibitory term of (8), the same kernels sample an LGN channel complementary to the one in the excitatory term. The net activity of model simple cells is thus a measure of image feature contrast in its preferred orientation. In mathematical terms, the kernels in (8) are:

\[
G_{pqijk}^{g+} = \frac{1}{2\pi \sigma_{lg} \sigma_{sg}} \exp \left\{-\frac{1}{2} \left\{ \frac{(p-i+m_k)\cos(\pi k/4) - (q-k+n_k)\sin(\pi k/4)}{\sigma_{lg}} \right.\right.
\]
\[
+ \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left. \left...
where \( Y_{ijk} \) is the activity described in (10), and the superscripts \( g^+ \) and \( g^- \) indicate opposite contrast polarities.

This computation yields twelve ‘hemispheric’ images, which are the outputs of boundary processing mechanisms via simple and complex cells in three spatial channels and four different orientations. To compute an unoriented boundary strength at each position and scale, we summed the boundary images across orientation:

\[
Z_{ij}^{g} = \sum_{k} z_{ijk}^{g}
\]

(12)

To prepare the data for ITa recognition, we flattened \( Z_{ij}^{g} \) from a two dimensional matrix to a row vector. That is, we took a matrix of \( M \) rows and \( N \) columns and mapped their values to an \( M \times N \) element row vector \( J \). Thus, for a given row \( m \) and column \( n \), that value would be mapped to element \( N(m-1)+n \) of \( J \); e.g. the value at row 2 column 2 in a 12 column matrix would map to element 14 of \( I \). The vector \( J \) is then normalized:

\[
I_i = \frac{J_i}{\max(J)}
\]

(13)

via an operation that emulates a shunting network with global inhibition, where \( J_i \in J \). This fixes all values of the normalized vector \( I \) within the interval \([0,1]\). Equations (1)-(13) are computed for all images, thus preparing three ‘flattened’ vectors (for three spatial scales) for each input image.

d) ITa Recognition

For the model ITa we used a variant of ART, Fuzzy ART\(^{31}\). Each input is the \( M \times N \)-dimensional ‘flattened’ vector \( I \). Each category \((j)\) corresponds to a vector \( w_j = (w_{j1},...,w_{jm}) \) of adaptive weights or long-term memory (LTM) traces. The number of potential categories \( N(j=1,...,N) \) is arbitrary. At first,

\[
w_{j1} = ...w_{jM} = 1
\]

(14)

and each category is then said to be uncommitted.

After a category is selected for coding it becomes committed. Each LTM trace \( w_{ji} \) is monotone nonincreasing through time and so converges to a limit. This property assures the stability of learned memories. In Fuzzy ART, both the bottom-up and top-down learned weights are the same, so a single weight vector \( w_j \) corresponding to each learned category suffices to represent both bottom-up and top-down learning. Dynamics are determined by a choice parameter \( a > 0 \), a learning rate parameter \( \beta \in[0,1] \), and a vigilance parameter \( \rho \in[0,1] \). For each input \( I \) and category \( j \), the bottom-up adaptive filter from the cell region representing
distributed features (V1 in the current simplified model) to the cell region representing learned
categories (ITa in the current simplified model) computes a choice function \( T_j \):

\[
T_j(I) = |I \land w_j| + (1 - \alpha)(M - |w_j|),
\]

where the fuzzy AND operator \( \land \) is defined by

\[
(x \land y)_i = \min(x_i, y_i)
\]

and where the norm \( |\cdot| \) is

\[
|x| = \sum_{i=1}^{M} |w_j|.
\]

One interpretation of the fuzzy AND operator is in terms of the fraction of learned postsynaptic
sites that can be activated by each input.

After all bottom-up inputs are registered at the category level, the cells compete via long-
range lateral inhibition to choose that category which receives the largest input, which is then
stored in short-term memory. For convenience, \( T_j(I) \) may be written as \( T_j \), and the category
choice is denoted by:

\[
T_j = \max \{T_j : j = 1...N\}.
\]

In the case of a tie, the cell with the smallest index is chosen. Selection of a category enables top-
down read-out of its learned expectation to the distributed feature level. As noted above, in fuzzy
ART, the same adaptive weights act in the top-down learned expectation as in the bottom-up
adaptive filter. Hence, attentive selection by biased competition uses the fuzzy AND operation
which, in particular, drives a cell’s response to zero if its top-down learned weight is zero.

As in Figure 1, each active bottom-up input tries to turn on the orienting system, and each
active cell at the distributed feature level tries to turn it off. In response to bottom-up activation
alone, before the category level gets activated, there are as many active features as inputs, so the
total inhibition \( |I| \) to the orienting system is sufficient to shut it off. When the top-down
expectation is active, however, it selects consistent features so that the total inhibition from the
remaining active features is reduced to \( |I \land w_j| \). Whether this amount of inhibition is sufficient
to prevent the orienting system from being activated depends upon the vigilance parameter \( \rho \),
because each excitatory input is multiplied by \( \rho \) to generate a total excitatory input to the
orienting system of \( \rho|I| \). Thus, vigilance determines the sensitivity of the orienting system to
bottom-up excitation.

If the total excitation is less than the total inhibition at the orienting system, then the
orienting system remains quiet and allows the bottom-up and top-down signals to cycle. That is
why resonance is said to occur if the match function of a chosen category meets the vigilance
criterion:
which just means that total inhibition is stronger than total excitation for that choice of $\rho$. Learning can then occur as described in equation (21) below; hence, the model name adaptive resonance. If, however, total excitation exceeds inhibition at the orienting system, then the orienting system can become active and generate a novelty-sensitive arousal burst that resets the currently active category. Mismatch reset is thus said to occur when

$$\frac{|I \land w_J|}{|I|} < \rho. \quad (20)$$

Then the value of the choice function $T_J$ is reset to -1 for the remainder of the input presentation to prevent its persistent selection during search. In more dynamical descriptions of ART, such persistent inhibition is accomplished by an interaction of habituative transmitters with the arousal burst, which together cause rebounds in cell activation (i.e., the -1) that are maintained by recurrent lateral inhibition among the category cells$^{20,33}$. The search process continues until (19) is satisfied, and then the weights are updated as follows:

$$w_{J}^{\text{new}} = \beta(I \land w_{J}^{\text{old}}) + (1 - \beta)w_{J}^{\text{old}}. \quad (21)$$

Fast learning can occur in fuzzy ART without causing catastrophic forgetting. Fast learning means that adaptive weights can reach their new equilibria on every learning trial. Our simulations were carried out under conditions of fast learning, for which $\beta = 1$ in (21). Also, for each simulation, we used three Fuzzy ART modules, one for each spatial scale.

To summarize, we implemented the following algorithmic instantiation of Fuzzy ART at three spatial scales:

1. Take the initial input from the current spatial scale $g$, $I_g^1$.
2. Initialize the weight vector to the initial category node, $w_1 = I_g^1$, and set the node to be committed.
3. Present the next input $I_n^g$.
4. Compute the activation to the category nodes via the choice-by-difference signal function$^{32}$, $T_j = |\tilde{T}_n^g \land w_j| + (1 - \alpha)(M - |w_j|)$, where $M$ is the number of dimensions in the input vector and $\alpha$ is the choice parameter.
5. Choose the category node $J$ with the largest input; that is, $J = \arg \max(\tilde{T}_j)$.
6. Check to see if the chosen category satisfies vigilance: $\frac{|\tilde{T}_n^g \land w_J|}{|I_n^g|} \geq \rho$.
   a. If vigilance is satisfied update the weights: $\tilde{w}_{J}^{\text{new}} = \beta(\tilde{T}_n^g \land w_{J}^{\text{old}}) + (1 - \beta)w_{J}^{\text{old}}$ where $\beta$ is the learning rate.
b. If vigilance is not satisfied, then set $T_j = -1$ and go to Step 5. If all nodes have been exhausted (which would never happen in vivo), add a new category node, set the weights to $\tilde{I}_n^g$ and set the node committed.

7. Go to Step 3 unless all inputs have been presented.

The parameters were set as follows: $\alpha = .0001$, $\rho = .1$, and $\beta = 1$. Readers seeking further detail should refer to the original Fuzzy ART paper \textsuperscript{31}. For a review of how the brain may use ART mechanisms to learn cortical recognition codes, see [34] and [35]. In particular, neurophysiological evidence for vigilance control in extrastriate cortex has been demonstrated in [36].

IV. Analysis

To measure a node’s selectivity, we utilized the same metric used by Zoccolan et al.\textsuperscript{5} on the responses of each node for the selectivity testing set:

$$S = \left[ 1 - \frac{\left( \sum R_i \right)^2}{n \left( \sum R_i^2 \right)} \right],$$ \hspace{1cm} (22)

where $R_i$ is the response of a category to input $i$ and $n$ is the total number of stimuli ($n = 203$). Then to measure tolerance, we took the mean of the inverse of the standard deviation of a node’s response for all objects across the three positions. The standard deviation gives an estimate of the ‘spread’ of the response, and thus the inverse would yield tolerance, since the higher the spread the lower the tolerance and vice-versa. Finally, we calculated the Pearson correlation coefficient between the two values as in Zoccolan et al.\textsuperscript{5}.

V. Detailed Results

Listed below are the Pearson correlation coefficients for each run for the four simulations.

<table>
<thead>
<tr>
<th></th>
<th>Run 1</th>
<th>Run 2</th>
<th>Run 3</th>
<th>Run 4</th>
<th>Run 5</th>
<th>Run 6</th>
<th>Run 7</th>
<th>Run 8</th>
<th>Run 9</th>
<th>Run 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sim. 1</td>
<td>-.359</td>
<td>-.589</td>
<td>-.467</td>
<td>-.419</td>
<td>-.279</td>
<td>-.63</td>
<td>-.311</td>
<td>-.449</td>
<td>-.312</td>
<td>-.279</td>
</tr>
<tr>
<td>Sim. 2</td>
<td>-.397</td>
<td>-.570</td>
<td>-.518</td>
<td>-.573</td>
<td>-.401</td>
<td>-.196</td>
<td>-.245</td>
<td>-.332</td>
<td>-.332</td>
<td>-.338</td>
</tr>
<tr>
<td>Sim .3</td>
<td>-.558</td>
<td>-.350</td>
<td>-.333</td>
<td>-.415</td>
<td>-.430</td>
<td>-.333</td>
<td>-.688</td>
<td>-.478</td>
<td>-.468</td>
<td>-.487</td>
</tr>
<tr>
<td>Sim. 4</td>
<td>-.602</td>
<td>-.598</td>
<td>-.143</td>
<td>-.422</td>
<td>-.280</td>
<td>-.669</td>
<td>-.390</td>
<td>-.258</td>
<td>-.482</td>
<td>-.367</td>
</tr>
</tbody>
</table>

To see how much of an effect the complex log transform had on the tradeoff, we ran two sets of simulations without it. This resulted in mean correlation coefficients of -.645 and -.676 with standard deviations of .083 and .06 for the first and second set, respectively. Hence, the
absence of the complex log significantly changed the degree of correlation without changing the
genral tradeoff.

<table>
<thead>
<tr>
<th></th>
<th>Run 1</th>
<th>Run 2</th>
<th>Run 3</th>
<th>Run 4</th>
<th>Run 5</th>
<th>Run 6</th>
<th>Run 7</th>
<th>Run 8</th>
<th>Run 9</th>
<th>Run 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sim. 1</td>
<td>-.756</td>
<td>-.571</td>
<td>-.738</td>
<td>-.585</td>
<td>-.466</td>
<td>-.643</td>
<td>-.666</td>
<td>-.689</td>
<td>-.632</td>
<td>-.705</td>
</tr>
<tr>
<td>Sim. 2</td>
<td>-.569</td>
<td>-.694</td>
<td>-.692</td>
<td>-.694</td>
<td>-.667</td>
<td>-.777</td>
<td>-.674</td>
<td>-.651</td>
<td>-.589</td>
<td>-.748</td>
</tr>
</tbody>
</table>

On the other hand, when using a single spatial scale, the tradeoff was nearly eliminated (results
not reported).

VI. Akrami et al. Simulation

We also simulated data from Akrami et al.\textsuperscript{15} on the response of neurophysiologically
recorded ITa cells to image morphs. Since they employed proprietary software to compute the
morphs, we simplified the procedure and used a progressive alpha blend from one image to the
other; that is, we started with just the first image in a pair visible and progressively made that
image more transparent and the second image less transparent until only the second image was
visible. As in the Zoccolan et al. simulations, we opted to employ natural images from the Cal
Tech 101 database. To train the model, we used 16 images for learning (see Figure S3) with the
parameters set as follows: $\alpha = 0.0001$, $\rho = 0.9$, and $\beta = 1$. Vigilance was chosen high because the
animals performed an active discrimination task; cf., [36]. Then, we froze learning and tested the
model on 8 morph pairs made from the training stimuli, similar to the procedure used by Akrami
et al. (2009).

Figure S3: Stimuli used to train the model for the Akrami et al. simulation.

VII. Implementation Details

The model was implemented entirely in Python using the NumPy\textsuperscript{37}, SciPy\textsuperscript{38} and
matplotlib\textsuperscript{39} libraries (sometimes referred to as PyLab). We see the general accessibility of
PyLab as a considerable strength. That is, the language is decidedly ‘readable’, and the entire
platform is free to the public. This, we hope, will lead to more widespread understanding and
scrutiny of the model. For possible extensions, the model was also implemented in C using
OpenCV.
References