Visual Categorization: How the Monkey Brain Does It

Ulf Knoblich[1], Maximilian Riesenhuber[1], David J. Freedman[2], Earl K. Miller[2], and Tomaso Poggio[1]

[1] Center for Biological and Computational Learning, McGovern Institute for Brain Research, Artificial Intelligence Lab and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
{knoblich,max,tp}@ai.mit.edu,
[2] Picower Center for Learning and Memory, RIKEN-MIT Neuroscience Research Center and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
davidf@mit.edu,ekm@ai.mit.edu

Abstract. Object categorization is a crucial cognitive ability. It has also received much attention in machine vision. However, the computational processes underlying object categorization in cortex are still poorly understood. In a recent experiment, Freedman *et al.* recorded from inferotemporal (IT) and prefrontal cortices (PFC) of monkeys performing a “cat / dog” categorization task ([1] and Freedman, Riesenhuber, Poggio, Miller, *Soc. Neurosci. Abs.*, 2001). In this paper we analyze the tuning properties of view-tuned units in our HMAX model of object recognition in cortex [2, 3] using the same paradigm and stimuli as in the experiment. We then compare the simulation results to the monkey neuron population data. We find that the view-tuned model units’ tuning properties are very similar to those of IT neurons observed in the experiment, suggesting that IT neurons in the experiment might respond primarily to shape. The population of experimental PFC neurons, on the other hand, shows tuning properties that cannot be explained just by stimulus tuning. These analyses are compatible with a model of object recognition in cortex [4] in which a population of shape-tuned neurons responding to individual exemplars provides a general basis for neurons tuned to different recognition tasks. Simulations further indicate that this strategy of first learning a general but object class-specific representation as input to a classifier simplifies the learning task. Indeed, the physiological data suggest that in the monkey brain, categorization is performed by PFC neurons performing a simple classification based on the thresholding of a linear sum of the inputs from examplar-tuned units. Such a strategy has various computational advantages, especially with respect to transfer across novel recognition tasks.

Visual Categorization:
How the Monkey Brain Does It

Abstract. Object categorization is a crucial cognitive ability. It has also received much attention in machine vision. However, the computational processes underlying object categorization in cortex are still poorly understood. In a recent experiment, Freedman *et al.* recorded from inferotemporal (IT) and prefrontal cortices (PFC) of monkeys performing a “cat / dog” categorization task ([1] and Freedman, Riesenhuber, Poggio, Miller, *Soc. Neurosci. Abs.*, 2001). In this paper we analyze the tuning properties of view-tuned units in our HMAX model of object recognition in cortex [2, 3] using the same paradigm and stimuli as in the experiment. We then compare the simulation results to the monkey neuron population data. We find that the view-tuned model units’ tuning properties are very similar to those of IT neurons observed in the experiment, suggesting that IT neurons in the experiment might respond primarily to shape. The population of experimental PFC neurons, on the other hand, shows tuning properties that cannot be explained just by stimulus tuning. These analyses are compatible with a model of object recognition in cortex [4] in which a population of shape-tuned neurons responding to individual exemplars provides a general basis for neurons tuned to different recognition tasks. Simulations further indicate that this strategy of first learning a general but object class-specific representation as input to a classifier simplifies the learning task. Indeed, the physiological data suggest that in the monkey brain, categorization is performed by PFC neurons performing a simple classification based on the thresholding of a linear sum of the inputs from examplar-tuned units. Such a strategy has various computational advantages, especially with respect to transfer across novel recognition tasks.

1 Introduction

The ability to group diverse items into meaningful categories (such as “predator” or “food”) is perhaps the most fundamental cognitive ability of humans and higher primates. Likewise, object categorization has received much attention in machine vision. However, relatively little is known about the computational architecture underlying categorization in cortex.

In [4], Riesenhuber and Poggio proposed a model of object recognition in cortex, HMAX, in which a general representation of objects in inferotemporal cortex (IT, the highest area in the cortical ventral visual stream, which is believed to mediate object recognition), provides the basis for different recognition tasks — such as identification and categorization — with task-related units located further downstream, *e.g.,* in prefrontal cortex (PFC). Freedman and Miller recently performed physiology experiments providing experimental population data for both PFC and IT of monkeys trained on a “cat / dog” categorization task ([1] and Freedman, Riesenhuber, Poggio, Miller, *Soc. Neurosci. Abs.*, 2001), allowing us to test this theory. In this paper, using the same stimuli as in the experiment, we analyze the properties of model IT units, trained without any explicit category information, and compare them to the tuning properties of experimental IT and PFC neurons. Compatible with the model prediction, we find that IT, but not PFC neurons show tuning properties that can be well explained by shape tuning alone. We then analyze the data to explore how a category signal in PFC could be obtained from shape-tuned neurons in IT, and what the computational advantages of such a scheme are.

2 Methods

2.1 The model

We used the HMAX model of object recognition in cortex [2, 3], shown schematically in Fig. 1. It consists of a hierarchy of layers with linear units performing template matching, and non-linear units performing a “MAX” operation. This MAX operation, selecting the maximum of a cell’s inputs and using it to drive the cell, is key to achieving invariance to translation, by pooling over afferents tuned to different positions, and scale, by pooling over afferents tuned to different scales. The template matching operation, on the other hand, increases feature specifity. A cascade of these two operations leads to C2 units (roughly corresponding to V4 / PIT neurons), which are tuned to complex features invariant to changes in position and scale. The outputs of these units provide the inputs to the view-tuned units (VTUs, corresponding to view-tuned neurons in IT [5, 3]). Importantly, the responses of a view-tuned model unit is completely determined by the shape of the unit’s preferred stimulus, with no explicit influence of category information.

2.2 Stimulus space

The stimulus space is spanned by six prototype objects, three “cats” and three “dogs” [1]. Our morphing software [7] allows us to generate 3D objects that are arbitrary combinations of the six prototypes. Each object is defined by a six-dimensional morph vector, with the value in each dimension corresponding to the relative proportion of one of the prototypes present in the object. The component sum of each object was constrained to be equal to one. An object was labeled a “cat” or “dog” depending on whether the sum over the “cat” prototypes in its morph vector was greater or smaller than those over the “dog” prototypes, resp. The class boundary was defined by the set of objects having morph vectors with equal cat and dog component sums.

2.3 Learning a population of cat/dog-tuned units

We performed simulations using a population of 144 VTUs, each tuned to a different stimulus from the cat / dog morph space. The 144 morphed animal stimuli were a subset of the stimuli used to train the monkey, *i. e.,* chosen at random from the cat / dog morph space, excluding “cats” (“dogs”) with a “dog” (“cat”) component greater than 40%. This population of VTUs was used to model a general stimulus representation consisting of neurons tuned to various shapes, which might then provide input to recognition task-specific neurons (such as for cat / dog categorization) in higher areas [4]. Each VTU had a tuning width of $\sigma = 0.2$ and was connected to the 32 C2 afferents that were most strongly activated by its respective preferred stimulus [2], which produced neurons with realistic broadness of tuning (see [8] for details).

Test set. The testing set used to determine an experimental neuron’s or model unit’s category tuning consisted of the nine lines through morph space connecting one prototype of each class. Each morph line was subdivided into 10 intervals, with the exclusion of the stimuli at the mid-points (which would lie right on the class boundary, with an undefined label), yielding a total of 78 stimuli.

2.4 Training a “categorization” unit (CU)

The activity patterns over a subset of VTU or C2 units to each of the 144 stimuli were used as inputs to train a Gaussian RBF output unit (see Fig. 1), which performed a weighted sum over its inputs, with weights chosen to have the CU’s output best match the stimulus’ class label (we used $+1$ for cat and -1 for dog as desired outputs, for details, see [9]). The performance of the categorization unit was then tested with the test stimuli described above (which were not part of the training set), and a classification was counted as correct if the sign of the CU matched the sign of the class label.

2.5 Evaluating category tuning

We use three measures to characterize the category-related behavior of experimental neurons and model units: the between-within index (BWI), the class coverage index (CCI) and the receiver operating characteristics (ROC).

BWI The *between-within index* (BWI) [1] is a measure for tuning at the class boundary relative to the class interior. Looking at the response of a unit to stimuli along one morph line, the response difference between two adjacent stimuli can be calculated. As there is no stimulus directly on the class boundary, we use 20% steps for calculating the response differences. Let btw be the mean response difference *between* the two categories (*i. e.,* between morph index 0.4 and 0.6) and wi the mean response difference *within* the categories. Then the between-within index is

$$\mathrm{BWI} = \frac{btw - wi}{btw + wi}. \tag{1}$$

Thus, the range of BWI values is -1 to $+1$. For a BWI of zero the unit shows on average no different behavior at the boundary compared to the class interiors. Positive BWI values indicate a significant response drop across the border (*e. g.,* for units differentiating between classes) whereas negative values are characteristic for units which show response variance within the classes but not across the boundary.

CCI The *class coverage index* (CCI) is the proportion of stimuli in the unit's preferred category that evoke responses higher than the maximum response to stimuli from the other category. Possible values range from $\frac{1}{39}$, meaning out of the 39 stimuli in the class only the maximum itself evokes a higher response than the maximum in the other class, to 1 for full class coverage, *i. e.,* perfect separability. When comparing model units and experimental neurons, CCI values were calculated using the 42 stimuli used in the experiment (see section 3.1), so the minimum CCI value was $\frac{1}{21}$.

ROC The *receiver operating characteristics* (ROC) curve [10] shows the categorization performance of a unit in terms of correctly categorized preferred-class stimuli (hits) *vs.* miscategorized stimuli from the other class (false alarms). The area under the ROC curve, A_{ROC}, is a measure of the quality of categorization. A value of 0.5 corresponds to chance performance, 1 means perfect separability, *i. e.,* perfect categorization performance.

3 Results

3.1 Comparison of model and experiment

We compared the tuning properties of model units to those of the IT and PFC neurons recorded from by Freedman [1] from two monkeys performing the cat / dog categorization task. In particular, the monkeys had to perform a delayed match-to-category task where the first stimulus was shown for 600ms, followed by a 1s delay and the second, test, stimulus. In the following, we restrict our analysis to the neurons that showed stimulus selectivity by an ANOVA ($p < 0.01$), over the 42 stimuli along the nine morph lines used in the experiment (in the experiment, stimuli were located at positions 0, 0.2, 0.4, 0.6, 0.8, and 1 along each morph line). Thus, we only analyzed those neurons that responded significantly differently to at least one of the stimuli.[1]

In particular, we analyzed a total of 116 stimulus-selective IT neurons during the "sample" period (100ms to 900ms after stimulus onset). Only a small number of IT neurons responded selectively during the delay period. For the PFC data, there were 67 stimulus-selective neurons during the sample period, and 32 stimulus-selective neurons during the immediately following "delay" period (300 to 1100 ms after stimulus offset).

Figs. 2 and 3 show the BWI, CCI, and A_{ROC} distributions for the IT neurons (during the sample period — IT neurons tended to show much less delay activity than the PFC neurons), and the PFC neurons (during the delay period — responses during the sample period tended to show very similar CCI and ROC values and slightly lower average BWI values (0.09 *vs.* 0.15).[2]

We compare the tuning of experimental neurons to that of the 144 model VTUs. As model units — unlike real neurons — show deterministic responses that might lead to artificially high ROC and CCI values, we chose to add independent Gaussian noise to the model unit responses for a fairer comparison (not adding any noise produces units with significantly higher CCI values and slightly higher ROC values; for more details, see [8]). We observe a very good agreement of simulation and experiment (Fig. 4): BWI and A_{ROC} distributions are not statistically significantly different ($p \geq 0.2$, Wilcoxon rank sum test), and the CCI distribution is only marginally different ($p = 0.06$).

3.2 Comparison of model units *vs.* PFC neurons

Unlike the IT neurons, the PFC neurons show a BWI distribution with a positive mean significantly different from zero (sample period: 0.09, delay: 0.15), combined with higher average CCI values (sample: 0.21, delay: 0.21), with single neurons reaching values as high as 0.76 (sample and delay). This maximum value lies outside the range of CCI values of model units. A positive average BWI of the magnitude found in the PFC data could only be obtained in the model with a significant number of of border-tuned neurons, but such border-tuned units have very low CCI values (see [8]). CCI values of PFC neurons are higher than those of IT neurons, however. Thus, the tuning properties of PFC

[1] Extending the analysis to include all *responsive* neurons (relative to baseline, $p < 0.01$) added mainly untuned neurons with CCIs close to 0, and A_{ROC} values close to 0.5.

[2] For comparison with the model, the indices and ROC curves were calculated using a neuron's averaged firing rate (over at least 10 stimulus presentations) to each stimulus.

neurons *cannot* be explained in the model by mere shape tuning alone, but seem to require the influence of explicit category information during training.

3.3 Categorization using C2 or VTU inputs

Unlike the VTUs whose training was *unsupervised*, *i. e.*, without any category information, the categorization unit (CU) was trained in a *supervised* fashion to respond with a positive value to cats and with a negative value to dogs. Fig. 5 shows the averaged categorization performance of CUs trained with either a randomly selected subset of C2 units or a randomly selected subset of the 144 VTUs as input.

We see that performance is substantially better for the CU trained on the VTU representation. For the same number of inputs, performance based on the C2 input representation generally lies below that obtained with a VTU representation. Quantitatively, the performance in particular of the VTU CU is similar to that of the monkeys on the same stimuli (whose performance was around 90%, [1]).

The advantage of using a VTU *vs.* a C2 input representation is likely due to the fact that the VTU are tuned to the objects the categorization task is defined over. This endows the system with good categorization performance already with few inputs. C2 units, on the other hand, do not show a selectivity that is particularly suitable for the task. This necessitates the use of a greater number of inputs for a CU based on C2 inputs to obtain a comparable level of performance.

4 Discussion

The different response properties of neurons in the two brain areas, with IT neurons coding for stimulus shape and PFC neurons showing more task-related tuning, are compatible with our recent model of object recognition in cortex [4, 3] in which a general object representation based on view- and object-tuned cells provides a basis for neurons tuned to specific object recognition tasks, such as categorization. This theory is also supported by data from another experiment in which different monkeys were trained on an identification and a categorization task, respectively, using the same stimuli [11], and which found no differences in the stimulus representation by inferotemporal neurons of the monkeys trained on different tasks. On the other hand, a recent experiment [12] reported IT neuron tuning emphasizing category-relevant features over non-relevant features (but no explicit representation of the class boundary, unlike in [1]) in monkeys trained to perform a categorization task. Further studies comparing IT neuron tuning before and after training on a categorization task or even different tasks involving the same set of stimuli, and studies that investigate the possibility of top-down modulation from higher areas (*e. g.*, PFC) during task execution, will be needed to more fully understand the role of top-down task-specific information in shaping IT neuron tuning.

In any case, the data and simulations support a very simple classifier, in which task-specific PFC neurons linearly combine inputs from IT, and category membership can be determined by a simple thresholding operation. The observed increase in the average CCI value from IT to PFC is compatible with this idea.

A further advantage of using an exemplar-based representation for categorization lies in the transfer across tasks such a representation affords: after learning a class-specific but task-independent representation for one task, it would be expected that learning another recognition task on the same set of objects would take less time than if the latter task were to be learned without the benefit of the former (note that this is a difference to other exemplar-based models of categorization [13, 14] where all categorization schemes have to be known at the time of learning, and all objects have to be represented in the same similarity space, making it difficult to represent situations where two objects can be judged as very similar under one categorization scheme but as rather different under another; as, for instance, a chili pepper and a candy apple in terms of color and taste, resp., see [9]). Such a behavior would be very attractive in human vision, where the same object usually belongs to a variety of categorization schemes (for instance, an apple can be categorized as “fruit”, or as “food” etc.), not all known in advance. It will be interesting to test this hypothesis in future experiments.

References

1. D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291:312–316, 2001.
2. M. Riesenhuber and T. Poggio. Are cortical models really bound by the “Binding Problem”? *Neuron*, 24:87–93, 1999.
3. M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2:1019–1025, 1999.
4. M. Riesenhuber and T. Poggio. Models of object recognition. *Nat. Neurosci. Supp.*, 3:1199–1204, 2000.
5. N.K. Logothetis, J. Pauls, and T. Poggio. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.*, 5:552–563, 1995.
6. K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cyb.*, 36:193–202, 1980.
7. C. Shelton. Three-Dimensional Correspondence. Master’s thesis, MIT, Cambridge, MA, 1996.
8. U. Knoblich, D.J. Freedman, and M. Riesenhuber. Categorization in IT and PFC: Model and Experiments. Technical Report AI Memo #2002-007, CBCL Paper #216, MIT AI Lab and CBCL, Cambridge, MA, 2002.
9. M. Riesenhuber and T. Poggio. A note on object class representation and categorical perception. Technical Report AI Memo 1679, CBCL Paper 183, MIT AI Lab and CBCL, Cambridge, MA, 1999.
10. N.A. Macmillan and C.D. Creelman. *Detection Theory: A User’s Guide.* 1991.

11. H.O. de Beeck, J. Wagemans, and R. Vogels. Inferotemporal neurons represent low-dimensional configurations of parametrized shapes. *Nat. Neurosci.*, 4:1244–1252, 2001.
12. N. Sigala and N. Logothetis. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415:318–320, 2002.
13. N. Intrator and S. Edelman. Learning low dimensional representations via the usage of multiple class labels. *NCNS*, 8:283–296, 1997.
14. S. Edelman. *Representation and Recognition in Vision*. MIT Press, Cambridge, MA, 1999.

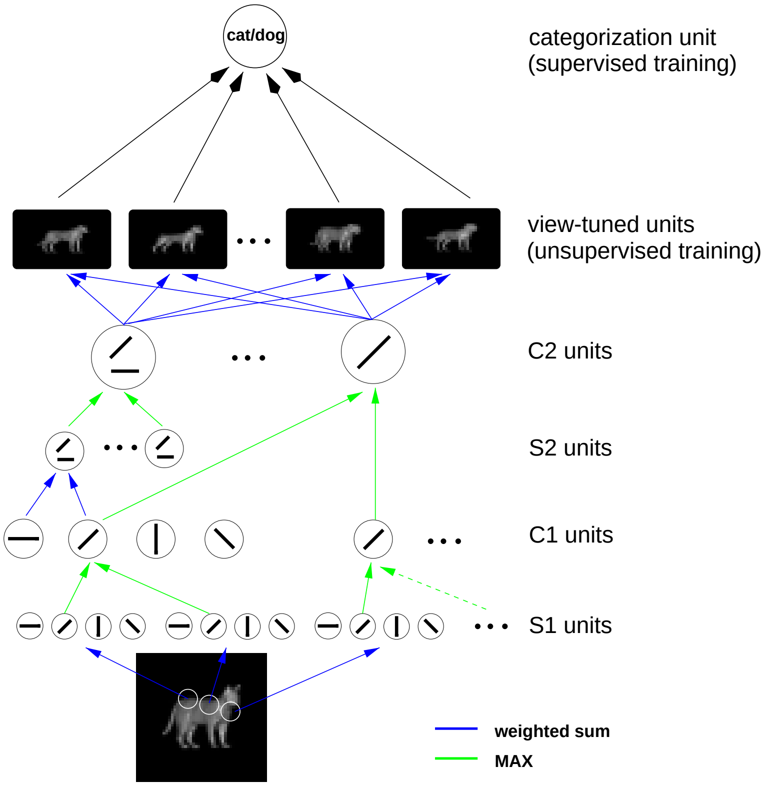

Fig. 1. Scheme of our model of object recognition in cortex [3]. The model consists of layers of linear units that perform a template match over their afferents ("S" layers, adopting the notation of Fukushima's Neocognitron [6]), and of non-linear units that perform a "MAX" operation over their inputs ("C" layers), where the output is determined by the strongest afferent (this novel transfer function is a key prediction of the model — for which there is now some experimental evidence in different areas of the ventral visual pathway (I. Lampl, M. Riesenhuber, T. Poggio, D. Ferster, *Soc. Neurosci. Abs.*, 2001; and T.J. Gawne and J.M. Martin, in press) — and a crucial difference to the aforementioned Neocognitron). While the former operation serves to increase feature complexity, the latter increases invariance by effectively scanning over afferents tuned to the same feature but at different positions (to increase translation invariance) or scale (to increase scale invariance, not shown).

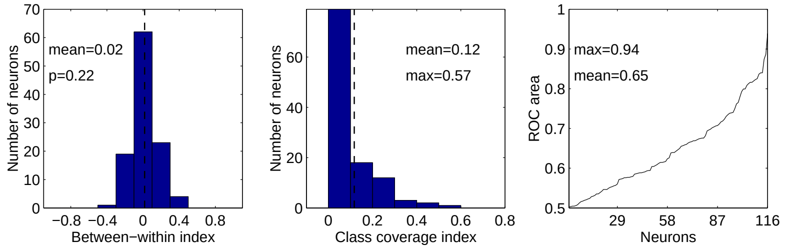

Fig. 2. Experimental IT data. The plots show the distribution of BWI (left), CCI (center) and ROC (right) area values.

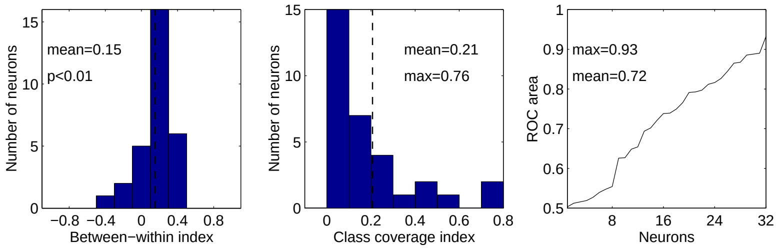

Fig. 3. Experimental PFC data (delay period). The plots show the distribution of BWI, CCI and ROC area.

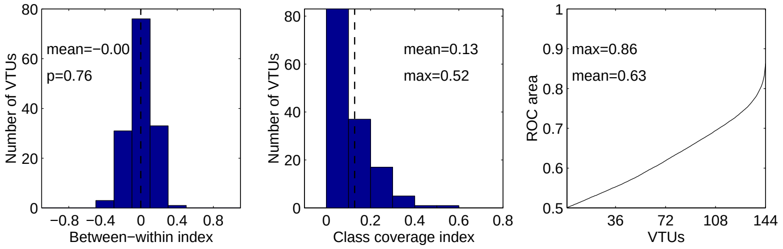

Fig. 4. Tuning of model VTUs with added independent Gaussian noise of amplitude $n = 0.1$. The plots show the distribution of BWI, CCI, and A_{ROC}. Values shown are the average over 100 trials.

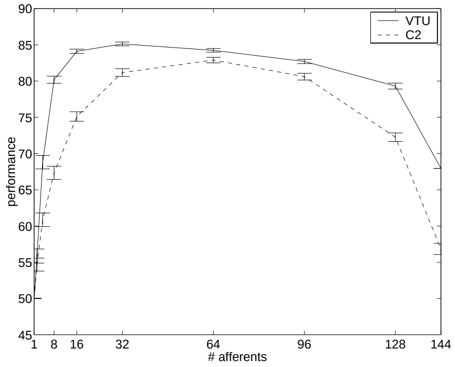

Fig. 5. Categorization performance of a "category neuron" trained on a stimulus space representation based on either different numbers of C2 units (selected at random from the 256 C2 units) or different numbers of cat / dog VTUs (chosen at random from the 144 cat / dog VTUs in the training set). Values shown are the average over 100 trials. Errorbars show the standard error of the mean. The decreasing performance for large numbers of units are due to overfitting.