Roles of the prefrontal neurons in delayed matching-to-category task: A modeling study

Tetsuto Minami * Toshio Inui

Department of Intelligence Science and Technology
Graduate School of Informatics
Kyoto University

Abstract

In understanding the environment around us, we do not perceive our exact sensory input as is. Instead, we divide the world into meaningful groups or categories. This categorization function is fundamental to cognitive processes. Such categories are processed in brain areas such as the inferior temporal cortex (ITC) and prefrontal cortex (PFC). To clarify the fundamental neuronal mechanisms involved in category information, we simulated the delayed matching-to-category task in Freedman et al. (2001) and showed that category selectivity in the PFC reflected the interaction between the selectivity of complex stimuli in the ITC and task-dependency in the motor area.

Key words: Perceptual categories, Recurrent network, Attractor dynamics, Prefrontal cortex, Working memory, Inferior temporal cortex

1 Introduction

In understanding the environment around us, we do not perceive our exact sensory input as is. Instead, we divide the world into meaningful groups or categories. This categorization function is fundamental to cognitive processes. For example, when we know that a new object belongs to a certain category, the categorical information provides us with its relevant parts and functions. Since categories often group together things that appear very different, their representation must involve something beyond physical appearance.

* Corresponding author
Email address: minami@cog.ist.i.kyoto-u.ac.jp (Tetsuto Minami).

The intricate behavioral skills of higher animals naturally depend on their categorization abilities. In addition, higher animals have an enormous ability to learn and adapt. Higher animals, such as monkeys, can acquire higher-level perceptual categories, such as food versus nonfood ([1]), tree versus non-tree, fish versus non-fish ([2]), and ordinal numbers ([3]).

Categories are processed in brain areas such as the inferior temporal cortex (ITC) and prefrontal cortex (PFC). The PFC and ITC are connected directly ([4,5]) and both areas include neurons that have selectivity to complex stimuli, such as trees, fish, faces, brushes, etc. ([6–8]). However, it is not clear whether the activities of such neurons reflect categories. The diagnostic characteristics of such neurons have not been tested, such as whether they have sharp boundaries or perform within-category analysis, so their activities might reflect only the similarities and differences in the physical appearance of stimuli, not necessarily their category membership.

To evaluate the role of the PFC in categorization, Freedman et al. (2001) taught monkeys to classify visual stimuli into two categories: dogs and cats. Stimuli were made by morphing original images of three prototype dogs and three prototype cats using computer graphics. Category membership was defined by whichever category contributed more ($> 50\%$) to a given morph. The monkeys performed a delayed matching-to-category (DMC) task that required them to judge whether two successive stimuli were from the same category. The results showed that the activity of many neurons sharply differentiated between the two categories that mirrored the monkeys' behavior. The neurons seem to encode the categories of the stimuli. They concluded that category information is represented in the PFC at the single-neuron level.

Knoblich et al. (2002) conducted the simulation of such category information processing. They simulated the physiological experiment of Freedman et al. (2001) using a model of object recognition. Their model consisted of a hierarchy of layers with linear units performing template matching, and non-linear units performing a "MAX" operation. Their model explained the category selectivity in the ITC, but their results were inconsistent with PFC neurons. In addition, they dealt with Freedman's task as a simple classification task, and did not consider the temporal dynamics involved.

The purpose of our study was to clarify the fundamental neuronal mechanisms for category information. We simulated the delayed matching-to-category task in Freedman et al. (2001) and showed that the category selectivity in the PFC reflected the interaction between the selectivity of complex stimuli in the ITC and task-dependency in the motor area.

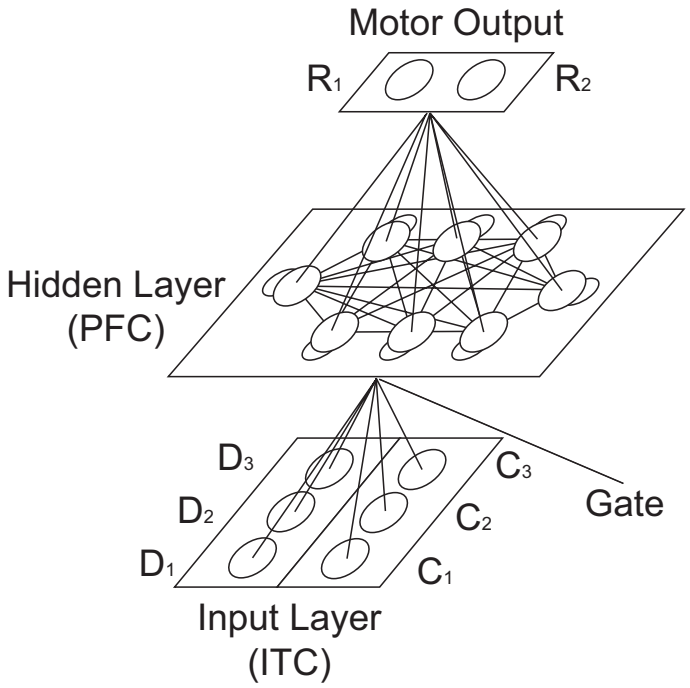


Fig. 1. The neural network architecture includes input, hidden, and output layers: The input layer consists of gate and object input. Each unit in the hidden layer receives connections from all of the units in the input layer. Hidden units also receive inputs from the rest of the hidden layer through recurrent connections. The hidden layer is fully connected to the output layer.

2 Modeling Methods

We proposed a neural network model for a delayed matching-to-category (DMC) task and simulated the results of physiological experiments. Our model was generated using a neural system identification technique. Neural system identification is interesting as a modeling paradigm because of the empirical finding that the internal behavior of neural identification models often mimics the internal behavior of the neural plant being modeled ([9]).

A schematic drawing of our neural network model is shown in Fig 1. It contains three layers: the input, hidden, and output layers. The input layer has two input lines: one input line for gate input and one for object input. The gate input carries a binary signal that is kept at zero as long as information is held in memory and set to 1.0 only when new information is gated. The object input represents the stimuli in a simple format, with separate units for the two different categories (DOG and CAT). There are three units within each category (DOG: D1, D2, D3, CAT: C1, C2, C3). The input lines connect to all of the model units, except the output units.

The network outputs consist of two units representing "response": $(1, 0)$ means

"release the lever" and $(0, 1)$ means "keep on holding the lever" (no response). The output units receive input from all of the hidden units, but do not feed back to them. The hidden layer consists of recurrently connected logistic units (30-60 units). In a recurrent network, the output of the network accounts for both the current inputs and the activities at earlier times. To compare the results with experimental recordings, one time step was considered to last 200 ms. The time lag of the recurrent connection was one time step.

The output of the ith model unit in the hidden layer at time $t + 1$ is given by

$$h_i(t + 1) = f(\sum_j hw_{ij} h_j(t) + \sum_k v_{ik} z_k(t) - b_i) \tag{1}$$

where hw_{ij} and v_{ik} represent recurrent connections from every hidden unit and weighted connections from every input unit z_k, respectively. Each unit has a bias, b_i. $f(x)$ is a logistic function

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

Each motor output at time $t + 2$ is defined as

$$o_m(t + 2) = f(\sum_i ow_{mi} h_i(t + 1) - b_m) \tag{3}$$

where ow_{mi} represents the weighted connections from the hidden units. Each unit has a bias, b_m.

The connections were optimized through a gradient descent using real-time recurrent learning (RTRL) ([10]). This algorithm computes the derivatives of states and outputs, with respect to all weights, as the network processes the sequence during the forward step. We used this algorithm to train the network to perform the required behavior, without claiming that the algorithm is similar to the learning mechanisms in the brain. Models were generated by training networks for roughly 10^7 network time steps. Training ceased when the mean square error was less than .01.

To train the model, randomized sequences of object and gate inputs were provided at appropriate times. The network output was trained to produce the response $(1, 0)$ when the category of the current object input matched that of the gated sample input and the response $(0, 1)$ when the category of the current object input did not match that of the gated sample input.

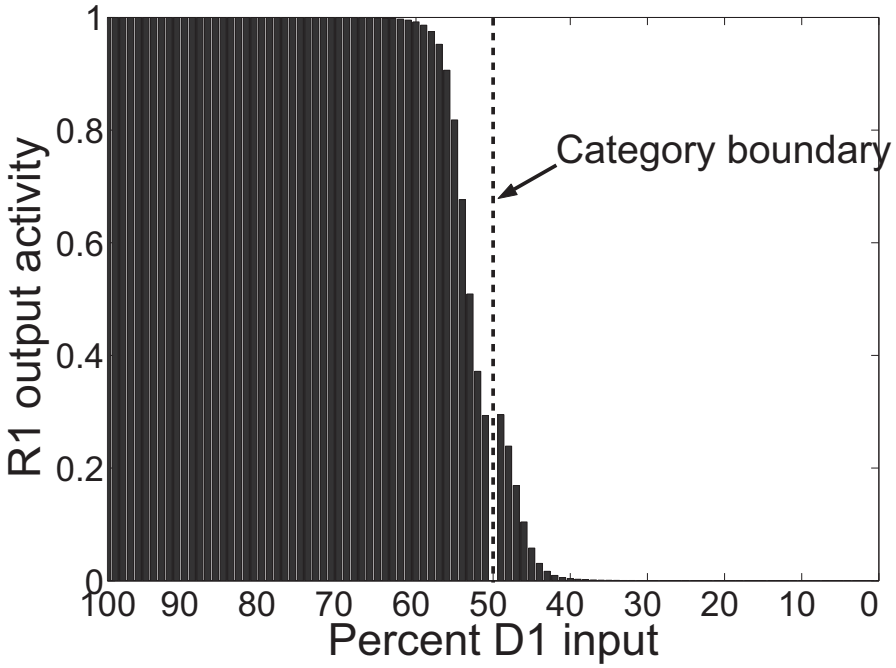


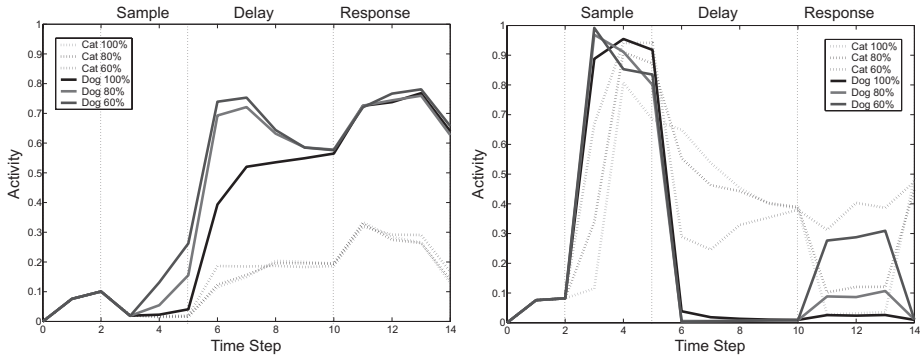
Fig. 2. Simulated psychophysical performance for the two-category tasks in the model.

3 Results

First, we trained the model to perform two-category tasks. After an inter-trial interval, an object input was gated in (steps 3 to 5). We called this period the sample period. After a delay (from steps 6 to 10), a test object was input from steps 11 to 13, which corresponded to the response period. The object input consisted of 54 patterns, which was the product of the combination of prototypes (9) and the ratio of DOG to CAT (6: 1.0, 0.8, 0.6, 0.4, 0.2, 0.0). We investigated the temporal pattern of the hidden units in a typical model network consisting of 40 hidden units.

Our model mimicked the psychophysical performance for the task (Fig. 2). As with the monkey's task performance, the network output did not decrease linearly as stimuli approached the category boundary, but changed more sharply at the category boundary.

Our model also reproduced the temporal patterns of the category-selective neurons in the PFC. The activity of many units in the hidden layer differed sharply between the two categories. That is, they showed relatively large differences in activity in response to samples from different categories and relatively similar activity in response to samples from the same category. Two examples of such units are shown in Fig. 3. Both seem to encode the category of the stimuli.



(a) DOG-category selective unit (b) CAT-category selective unit

Fig. 3. Examples of category units (two-category task): (a) the activity of a unit that showed greater activity in response to the category DOG during the delay period. The histogram traces represent the unit's activity in response to stimuli at each of the six category levels. (b) The activity of a unit that showed greater activity in response to the category CAT during the delay period.

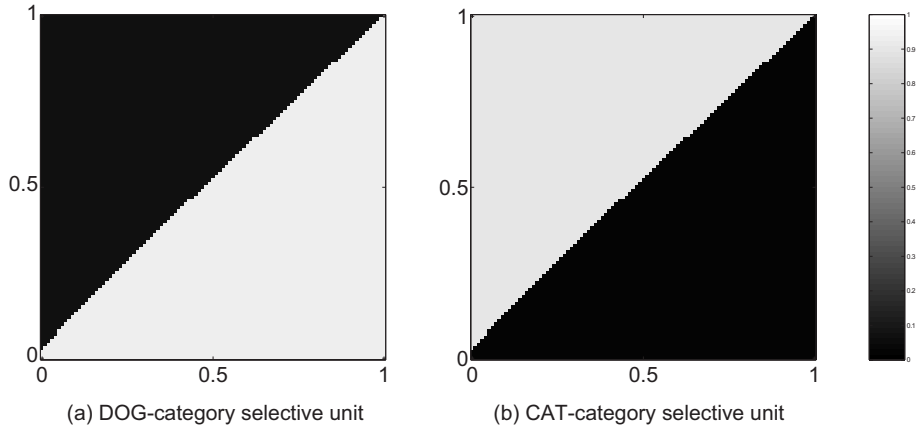


Fig. 4. Map of the category boundary between two basins of attraction: Steady state activation is plotted as a function of the gated (D_1, C_1) category stimulus across a 101 $\times$ 101 matrix. Here, (a) is a DOG-selective unit and (b) is a CAT-selective unit.

Next, we investigated the stable states of our model to clarify the mechanism of category classification. The stable states of our model are fixed point attractors. To plot the basins of attraction, we systematically gated the results in a 101 $\times$ 101 grid covering the two-dimensional category input space (Fig. 4). The figure shows that our model stored the remembered category in two basins of attraction.

What is the representation of category membership in our hidden layer? To quantify the effect of category membership, we computed a category index (CI) that reflected the difference in activity in response to samples across the category boundary versus the difference in response to samples from the same

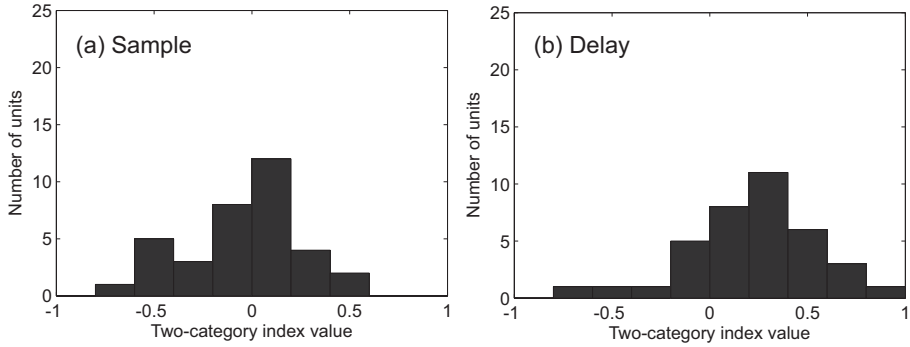


Fig. 5. Distribution of two-category index values across the entire population of 40 units during the (a) sample and (b) delay periods.

category. The CI equals $\frac{BCD-WCD}{BCD+WCD}$, in which BCD is the average difference of samples from different categories and WCD is the average difference of samples from the same category. Fig 5 shows the distribution of the two-category index values across the entire population of 40 units during the (a) sample and (b) delay periods. The index ranges from -1.0 to 1.0. Positive values indicate larger differences in neuronal firing between categories than within categories. Negative values indicate larger differences within categories than between categories. As the figure shows, the category indices were significantly greater (shifted more toward category tuning) during the delay than during the sample period, which is consistent with the physiological results.

Hitherto, we have examined the selectivity during the sample and delay periods. Next, we investigated the response property during the response period. During the response period, the model must decide whether a category matches that of the sample. We found units that showed match/nonmatch effects that were limited to one of the categories, while no units that showed an effect of match/nonmatch regardless of the category were found in the hidden layer. In our model, such units correspond to the output units.

In addition, we also simulated a three-category task. We retrained the network to implement the three-category task. Our model reproduced the temporal patterns of single neurons recorded during the three-category task. The units in the hidden layer showed selectivity to the three categories. By contrast, when the data were sorted according to the old CAT and DOG categories, there was no selectivity to the old categories. Therefore, our simulation simulated the effects of learning category representations in the physiological experiments.

These considerations and the analysis of connection weights in the network should help to elucidate the mechanism involved in performing a DMC task. We proposed an outline of the processing network involved in the DMC task. In it, selective activities in response to complex stimuli in the input layer corresponding to the ITC were processed into category-selective activities dur-

ing the interval between the sample period and the delay period. During the response period, the model determined whether the input was match or nonmatch after processing match/nonmatch limited to each category.

4 Conclusion

To explore the role of the PFC in categorization, we proposed a neural network model for a delayed matching-to-category (DMC) task and simulated physiological data for monkeys that were trained to perform the DMC task([11–13]). We showed that the category-selective activities in the PFC reflect the process of discriminating the input from the ITC according to the task. In addition, our model reproduced the essential physiological results without making any unnatural assumptions. Therefore, category selectivity in the PFC can be explained using bottom-up processes from the ITC to the PFC. With respect to the category selectivity of ITC neurons, it is likely that their selective responses reflect a top-down signal from the PFC.

Acknowledgments

This research was supported as part of the Neuroinformatics Research in Vision Project of the Advanced and Innovational Research program in Life Sciences, funded by Special Coordination Funds for Promoting Science and Technology, and a Grant-in-Aid for JSPS Fellows, 14002026, both from the Ministry of Education, Culture, Sports, Science, and Technology, of Japan.

References

[1] M. Fabre-Thorpe, G. Richard, S. J. Thorpe, Rapid categorization of natural images by rhesus monkeys., Neuroreport 9 (2) (1998) 303–8.

[2] R. Vogels, Categorization of complex visual images by rhesus monkeys. part 1: behavioural study., Eur J Neurosci 11 (4) (1999) 1223–38.

[3] T. Orlov, V. Yakovlev, S. Hochstein, E. Zohary, Macaque monkeys categorize images by their ordinal number., Nature 404 (6773) (2000) 77–80.

[4] L. G. Ungerleider, D. Gaffan, V. S. Pelak, Projections from inferior temporal cortex to prefrontal cortex via the uncinate fascicle in rhesus monkeys., Exp Brain Res 76 (3) (1989) 473–84.

[5] M. J. Webster, L. G. Ungerleider, J. Bachevalier, Connections of inferior temporal areas te and teo with medial temporal-lobe structures in infant and adult monkeys., J Neurosci 11 (4) (1991) 1095–116.

[6] R. Desimone, T. D. Albright, C. G. Gross, C. Bruce, Stimulus-selective properties of inferior temporal neurons in the macaque., J Neurosci 4 (8) (1984) 2051–62.

[7] E. K. Miller, C. A. Erickson, R. Desimone, Neural mechanisms of visual working memory in prefrontal cortex of the macaque., J Neurosci 16 (16) (1996) 5154–67.

[8] S. P. Scalaidhe, F. A. Wilson, P. S. Goldman-Rakic, Face-selective neurons during passive viewing and working memory performance of rhesus monkeys: evidence for intrinsic specialization of neuronal coding., Cereb Cortex 9 (5) (1999) 459–75.

[9] D. Zipser, Identification models of the nervous system., Neuroscience 47 (4) (1992) 853–62.

[10] R. J. Williams, D. Zipser, A learning algorithm for continually runnfing fully recurrent neural networks., Neural Computation 1 (1989) 270–280.

[11] D. J. Freedman, M. Riesenhuber, T. Poggio, E. K. Miller, A comparison of primate prefrontal and itemporal cortices during visual categorization., J Neurosci 23 (12) (2003) 5235–46.

[12] D. J. Freedman, M. Riesenhuber, T. Poggio, E. K. Miller, Visual categorization and the primate prefrontal cortex: neurophysiology and behavior., J Neurophysiol 88 (2) (2002) 929–41.

[13] D. J. Freedman, M. Riesenhuber, T. Poggio, E. K. Miller, Categorical representation of visual stimuli in the primate prefrontal cortex., Science 291 (5502) (2001) 312–6.