A MODEL OF IMPLICIT ASSOCIATION LEARNING BASED ON PLASTICITY IN THE PERIRHINAL CORTEX

Shigemitsu MOROKAMI[*], *Atsuo SUEMITSU*[**], *Masahiko MORITA*[***]

[*]Doctoral Program in System Information Engineering, University of Tsukuba
[**]Doctoral Program in Engineering, University of Tsukuba
[***]Institute of Engineering Mechanics and Systems, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
E-mail: morokami@bcl.esys.tsukuba.ac.jp

ABSTRACT

In the delayed match-to-sample task, responses of inferior temporal neurons to adjacent stimuli in the sequence are correlated to each other when the monkey was trained repeatedly with the sequence of visual stimuli, although the monkey was not required to associate the stimuli with each other. This correlation, however, is not observed for a monkey with lesions of the rhinal cortex, which is not consistently explained by existing models of such correlated responses. In the present study, we construct a model consisting of two networks corresponding to area TE and the perirhinal cortex, and show that perirhinal plasticity may underlie the mechanism of implicit association learning.

1. INTRODUCTION

It is known that the inferior temporal cortex (IT) plays a major role not only in visual recognition but also in visual short-term and long-term memories. Among the many examples of physiological evidences supporting this, the data on memory-related neurons of the monkey reported by Miyashita[1] is one of the most important. This data reveals the phenomenon that after a monkey was trained repeatedly on a delayed match-to-sample (DMS) task with a fixed sequence of visual stimuli, the neuronal responses to adjacent stimuli in the sequence are correlated to each other, although the monkey was not required to associate the stimuli explicitly. This is very interesting from the viewpoint of the representation and structuralization of long-term memory in the brain.

To explain this phenomenon, two theoretical models have been presented; however, they do not match the physiological evidence in a critical point. In particular, they both have difficulty in explaining the result of the lesion study on the rhinal cortex (perirhinal and entorhinal cortices) which plays an important role in the formation process of visual memory.

In this paper, we present a model based on the findings on the perirhinal cortex, and show that the association between adjacent stimuli is quite possibly learned by a mechanism different from those mentioned above.

2. BACKGROUND

2.1. IT neurons related to implicit association

The empirical study by Miyashita [1] is summarized as follows.

First, he trained two monkeys repeatedly on a DMS task for two weeks using 97 fractal pictures generated by a computer. In this task, a picture is presented for a short time as a sample stimulus, and the monkey must judge whether a picture (test stimulus) presented after a delay of 16 seconds is the same as the sample or not. The pictures were numbered from 1 to 97 and used as a sample always in that order.

Then he recorded the neuronal activities in IT (mainly in area TE) while the trained monkeys were performing the task with the pictures used in training (learned set) and with 97 novel pictures (unlearned set), and found that many neurons exhibit strong activity during the delay period after the sample presentation of some specific learned pictures. The pictures eliciting a strong response to each of these neurons did not particularly have common features or similarity, but they were often near one another in the picture sequence in training. That is, a neuron strongly responding to a learned picture tends to show a large response to neighboring pictures in the learned set.

This result is important in that the delay activities reflect the temporal relation between stimuli in the learning process, which clearly differs from selective response to pictorial features as is usually seen in IT. It should also be noted that the monkey has only to remember the sample picture during a single trial to perform the DMS task used in this experiment, and

need not associate pictures with each other. In this sense, such association between neighboring pictures is formed implicitly, which we will refer to as implicit association.

A similar phenomenon has been observed by Sakai and Miyashita [2] in the case of a pair-association (PA) task, in which the monkey must associate a pair of visual stimuli explicitly. They reported that neurons responding to both of the paired pictures ('pair-coding' neurons) were significantly more neurons than those responding to two unrelated pictures. Since in training, paired pictures were presented sequentially with a delay interval, we can regard such neuronal activity as reflecting the temporal closeness of the visual stimuli.

2.2. Existing theories

At present, there exist two theoretical models explaining the mechanism of implicit association: the attractor model by Griniasty et al. [3] and Brunel [4], and the layer model by Wallis [5].

The former uses a recurrent neural network as a model of the IT network, where the state encoding a learned picture is an attractor of the network and the activity pattern ('code') representing the sample is maintained not only during the delay period but also during the intertrial interval. That is, information on a sample picture is preserved until sample presentation of the next picture, which enables the model to correlate the two picture codes.

On the other hand, the latter model uses a simple layered network with the trace rule, which is an extended Hebb rule, where information on the previous picture is preserved postsynaptically in the form of a trace value that is elevated by excitation of the presynaptic neuron or the input signal and then decays gradually. Since the synapse is reinforced if the postsynaptic neuron is excited while some trace value remains, the response pattern to a picture becomes similar to the code of the previous picture.

However, neither mechanism of intertrial memory retention in the two models has been physiologically verified or seems plausible. Moreover, these models do not well explain some important findings concerning the rhinal cortex.

2.3. Rhinal cortex

The rhinal cortex is an area composed of the perirhinal cortex (PRh), which is part of IT and adjacent to TE, and the entorhinal cortex which is in the medial temporal lobe. This area, particularly PRh, has recently been drawing attention for its role in memory, and many physiological findings have been obtained; we briefly describe some of them (see [6] for details).

First, lesions of PRh of the monkey moderately impair the learning and performance of DMS tasks, whereas simple visual recognition is not damaged. In contrast, they cause critical impairments in association learning; for example, a monkey with rhinal lesions is completely incapable of learning a new set of stimuli in a PA task.

Second, TE neurons of a monkey in the hemisphere, from which the rhinal cortex had been removed before training, show the same stimulus selectivity as those in the intact hemisphere, but do not show the 'pair-coding effect' [7], i.e., the high correlation between the responses to paired pictures.

These findings indicate that the rhinal cortex is critical for association learning of visual stimuli, which does not accord well with either of the above models in which the rhinal cortex is not taken into consideration and information on the previous sample is retained at the neurons exhibiting picture-selective delay activity; according to those models, although they deal with the implicit learning in DMS tasks, the pair-coding effect in TE will not disappear even after the removal of the rhinal cortex because paired pictures in the PA task are presented close together in time.

In connection with this, it is known that the rhinal cortex, particularly PRh, has high plasticity in two respects. One is synaptic plasticity; for example, Tokuyama et al. [8] demonstrated that the brain-derived neurotrophic factor was upregulated specifically in PRh during PA learning, indicating that the synaptic plasticity in PRh is related to associating visual stimuli.

The other is the phenomenon that neurons exhibit a reduced activity during the second or subsequent exposure to a stimulus relative to the first [9], which is called stimulus-specific adaptation (SSA) or the repetition inhibition effect. This adaptation effect lasts for a long time (more than a few minutes), and is not greatly affected by intervening stimuli presented between the first presentation and the second presentation of the stimulus.

Among the neurons exhibiting SSA, those in which the response varies according to not the familiarity but by the recency of the stimulus are referred to as recency neurons. Although recency neurons are also found in TE and the entorhinal cortex, their ratio is particularly large in PRh, and the duration of adaptation of recency neurons is significantly longer in PRh than in TE. These findings suggest that the information on a sample picture is retained by the recency neurons in

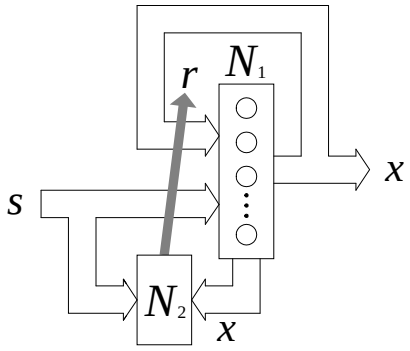

Figure 1: Block diagram of the model.

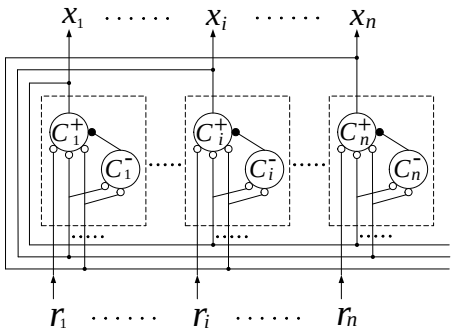

Figure 2: Structure of the association network.

PRh to form implicit association.

3. THE MODEL OF IMPLICIT LEARNING FORMATION

We previously constructed a model composed of two neural networks corresponding to TE and PRh, and have shown that it agrees well with the activities of IT neurons during a PA task [10]. This model, however, is for explicit association and does not explain the implicit association phenomenon.

Taking account of the above physiological findings, we modify this model by introducing synaptic plasticity and 'recency cells' showing the adaptation effect in the network corresponding to PRh. The structure of the model is shown in Fig.1, where association network N_1 and trainer network N_2 are interconnected.

3.1. The association network

The association network N_1 consists of pairs of excitatory and inhibitory cells (Fig.2). The excitatory cell C_i^+ receives a signal r_i from N_2 with input intensity of λ and recurrent inputs from other units, and emits output of the unit x_i. The inhibitory cell C_i^- sends a strong inhibitory signal to C_i^+. In mathematical terms,

$$y_i = f\left(\sum_{j=1}^{n} w_{ij}^- x_j - \theta\right), \quad (1)$$

$$\tau \frac{du_i}{dt} = -u_i + \sum_{j=1}^{n} w_{ij}^+ x_j - w^* y_i + z_i, \quad (2)$$

$$x_i = f(u_i), \quad (3)$$

where w_{ij}^+ and w_{ij}^- are the synaptic weights from the jth unit to C_i^+ and C_i^-, w^* represents the efficiency of the inhibitory synapse from C_i^+ to C_i^-, and θ is a threshold. The activation function $f(u)$ of each cell is a monotonic sigmoid function increasing from 0 to 1 given by

$$f(u) = \frac{1}{1 + e^{-cu}}. \quad (4)$$

In parallel with this, using a learning signal generated by N_2, learning of the synaptic weights of N_1 is performed according to

$$\tau' \frac{dw_{ij}^+}{dt} = -w_{ij}^+ + \alpha_1 r_i x_j, \quad (5)$$

$$\tau' \frac{dw_{ij}^-}{dt} = -w_{ij}^- - \beta_1 r_i x_j + \beta_2 x_i x_j + \gamma, \quad (6)$$

where α_1, β_1 and β_2 are learning coefficients, γ is a positive constant representing lateral inhibition among units, and τ' is a time constant of learning ($\tau' \gg \tau$).

3.2. The trainer network

The structure of the trainer network (N_2) is shown in Fig.3. This network consists of n pairs of output cell C_i^o and recency cell C_i^r which inhibit C_i^o, corresponding to the n units of N_1, and transforms the input pattern $\boldsymbol{s} = (s_1, ..., s_n)$ into the learning signal $\boldsymbol{r} = (r_1, ..., r_n)$ for the association network.

The ith cell C_i^o receives the input pattern $\boldsymbol{s} = (s_1, ..., s_n)$ and the feedback signal x_j from N_1 through synaptic weights p_{ij} and q_{ij}, respectively, and emits r_i to the ith unit of N_1. The output r_i is also sent to the recency cell C_i^r, which sends an inhibitory signal z_i back to C_i^o, and transiently increases the threshold or the degree e_i of fatigue depending on its value (we model SSA simply using the fatigue of cells, since the mechanism of SSA is not clear and modeling it for itself is not our purpose). Accordingly if C_i^o is strongly

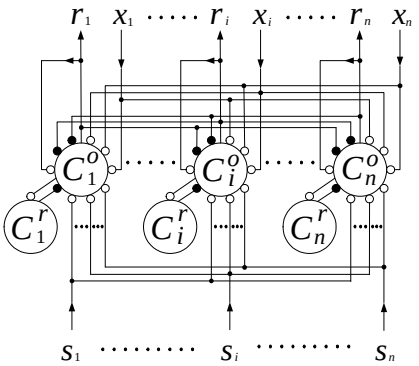

Figure 3: Structure of the trainer network.

activated, the inhibitory signal from C_i^r decreases for a while, when C_i^o is readily excited.

In mathematical terms,

$$\tau \frac{dv_i}{dt} = -v_i + \sum_{j=1}^{n} p_{ij} s_j + \sum_{j=1}^{n} q_{ij} x_j - \rho \sum_{j \neq i} r_j + \sigma r_i - \eta r_i^*, \quad (7)$$

$$r_i = f(v_i), \quad (8)$$

$$r_i^* = (\zeta - e_i) r_i, \quad (9)$$

$$\tau'' \frac{de_i}{dt} = -e_i + \iota r_i, \quad (10)$$

where v_i denotes the potential of C_i^o, ρ and σ represent the efficiency of lateral inhibition and self-excitation, respectively, η is the input weight from the recency cell, and ζ and ι are positive constants.

This network generates the learning signal as follows. First, when N_2 receives an external input pattern, say A, it emits a pattern a encoding A, and N_1 is trained so that a state close (in the sense of the vector direction) to the current learning signal $r =$ a can be a stable attractor. Then if N_2 receives a pattern B in the next trial, it emits a different pattern b that was a small correlation with a because of the adaptation effect of the recency cells. Since this effect decreases with time and does not disappear immediately, the output patterns c, d, and so forth of N_2 corresponding to the input patterns C, D, and so forth have some decreasing degree of correlation with a.

At the same time, the synaptic weights p_{ij} are modified according to

$$\tau' \frac{dp_{ij}}{dt} = -p_{ij} + \alpha_2 r_i s_j, \quad (11)$$

where α_2 is a positive constant, which makes the input signal to N_2 similar to the current learning signal. Because of this effect, the degree of correlation between a and b gradually increases to a certain value by repeating the same learning process.

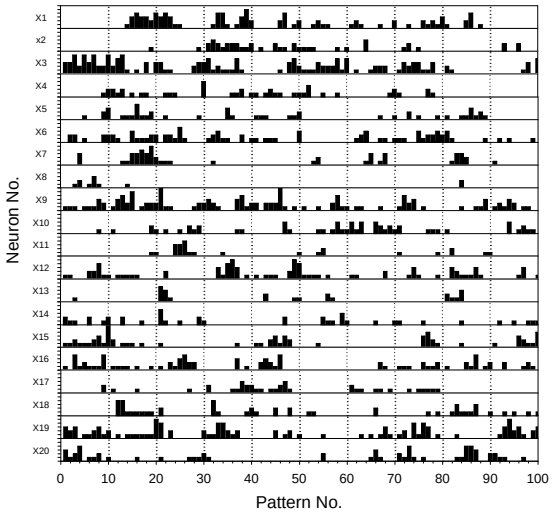

Figure 4: Responses of the units to each input pattern.

4. COMPUTER SIMULATIONS

We carried out computer simulations on the model with 1000 units. First we prepared 100 patterns that were 1000-dimensional vectors with 10% of the elements being 1 and the rest 0, and trained the model by feeding the patterns one by one for 4τ for each pattern. The parameters were

$\tau' = 50000\tau$, $\tau'' = 60\tau$, $\theta = 3.0$,$w^* = 10$,
$\alpha_1 = 50$, $\alpha_2 = 0.5$, $\beta_1 = 25$, $\beta_2 = 50$,
$\gamma = 0.05$, $\zeta = 0.9$, $\iota = 0.8$, $c = 10$.

After training, we examined the response of the units of N_1. Fig.4 displays the responses of some units to the 100 learned patterns, where units with various types of pattern selectivity can be seen; that is, some units respond strongly to only a few patterns and moderately to their neighboring patterns, and others show the same level of response to a long sequence of patterns. It should be noted that neurons of both types of response are actually observed in monkey IT.

We analyzed this simulation data in the same way as in Miyashita's study [1]. The result is shown in Fig.5, where correlation between the response to an input pattern and that to neighboring patterns is plotted by the solid line; the broken line represents the IT neurons (adapted from [1]). We see that the simulation result agrees well with the empirical data.

Finally, to examine the learning process of implicit association, we calculated the correlation between re-

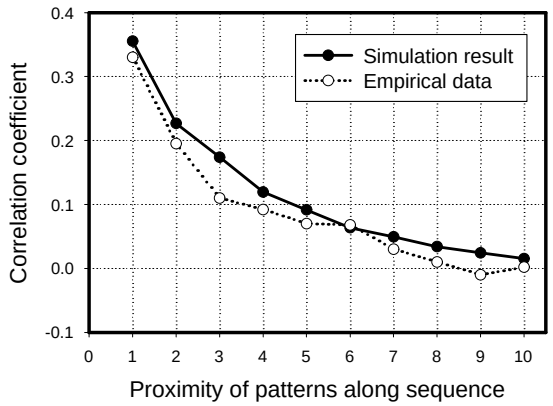

Figure 5: Correlation coefficient between responses to a pair of learned patterns versus their distance in the pattern sequence.

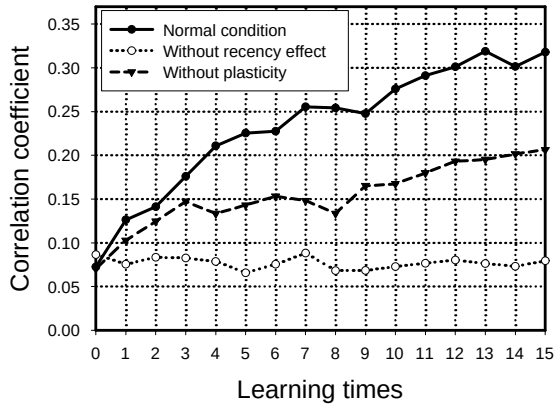

Figure 6: Learning process of implicit association.

sponses to adjacent patterns for every cycle of training. For comparison, we repeated simulations in the cases that $e_i \equiv 0$, or the recency effect was removed, and in the case that N_2 has no synaptic plasticity. Fig.6 shows the result, where the solid line indicates the intact model, and the broken and dotted lines indicate the cases without the recency effect and without synaptic learning of N_2, respectively. As is evident from this figure, no implicit association phenomenon was observed without the recency effect cells; even with the recency effect, if N_2 does not have synaptic plasticity, the degree of correlation does not increase much. This suggests that both the recency effect and synaptic learning in PRh are important for implicit association learning in TE.

5. CONCLUDING REMARKS

Based on the findings on the rhinal cortex, we have constructed a neural network model that forms implicit association, and have shown that it agrees well with the empirical data on IT neurons. This model is superior to other theoretical models in that it matches physiological evidence more closely, particularly the fact that the implicit association phenomenon in TE disappears when the rhinal cortex is removed. Another advantage is that this model can learn and perform a PA task and thus gives a unified theory of implicit and explicit association learning processes.

Empirical verification of the model remains for future study as well as the development of a model for recency neurons based on a more plausible mechanism.

6. REFERENCES

[1] Y. Miyashita, "Neural correlate of visual associative long-term memory in the primate temporal cortex," Nature, vol.335, pp.817–820, Oct. 1988.

[2] K. Sakai and Y. Miyashita, "Neural organization for the long-term memory or paired association," Nature, vol.354, pp.152–155, Nov. 1991.

[3] M. Griniasty, M.V. Tsodyks and D.J. Amit, "Conversion of the temporal correlations between stimuli to spatial correlations between attractors," Neural Computation, vol.5, pp.1–17, 1993.

[4] N. Brunel, "Hebbian learning of context in recurrent neural networks," Neural Computation, vol.8, pp.1677–1710, 1996.

[5] G. Wallis, "Spatio-temporal influences at the neural level of object recognition," Network: Computation in Neural Systems, vol.9, pp.265–278, 1998.

[6] E.A. Murray and T.J. Bussey, "Perceptual-mnemonic functions of the perirhinal cortex," Trends in Cognitive Sciences, Vol.3, pp.142–151, April. 1999.

[7] S. Higuchi and Y. Miyashita, "Formation of mnemonic neural response to visual paired associates in inferotemporal cortex is impaired by perirhinal and entorhinal lesions," Proc. Natl. Acad. Sci. USA, vol.93, pp.739–743, Jan. 1996.

[8] W. Tokuyama, H. Okuno, T. Hashimoto, Y.X. Li and Y. Miyashita, "BDNF upregulation during declarative memory formation in monkey inferior temporal cortex," Nature Neuroscience, vol.3, no.11, pp.1134–1142, Nov. 2000.

[9] J.L. Ringo, "Stimulus specific adaptation in inferior temporal and medial temporal cortex of the monkey," Behavioural Brain Research, vol.76, pp.191–197, 1996.

[10] M. Morita and A. Suemitsu, "Computational modeling of pair-association memory in inferior temporal cortex," Cognitive Brain Research, vol.13, pp.169–178, 2002.