A neural network model of pair-association memory in the inferotemporal cortex

Atsuo SUEMITSU, Masahiko MORITA***

*Doctoral Program in Engineering, University of Tsukuba
1-1-1 Ten-nodai, Tsukuba, Ibaraki 305-8573, Japan
**Institute of Engineering Mechanics and Systems, University of Tsukuba
1-1-1 Ten-nodai, Tsukuba, Ibaraki 305-8573, Japan

Email:sue@bcl.esys.tsukuba.ac.jp

Abstract

Neurons related to pair-association memory have been found in the inferotemporal cortex of monkeys, but their activities do not accord with existing neural network models. The present paper describes a neural network model consisting of excitatory-inhibitory cell pairs, which recalls paired patterns based on a gradual shift of the network state. It is demonstrated by computer simulations that this model agrees well with the observed neuronal activities.

1. Introduction

In the inferotemporal cortex of monkeys, interesting neuronal activities related to pair-association memory have been reported [1]. This finding is very important in considering how long-term memories are structured and retrieved in the brain.

This empirical data, however, cannot be explained by existing neural network models, since conventional dynamics of artificial neural networks do not accord with a gradual change in sustained neuronal activities.

In the present paper, we construct a first-step model of the pair-association memory consistent with the above finding.

2. Pair-association neurons in the monkey inferotemporal cortex

Sakai and Miyashita [1] studied inferotemporal neurons of monkeys by the following experiment.

They first generated many pairs of figures, and trained monkeys to associate paired figures with each other. Then they measured neuronal activities in a delayed matching task, where one of the paired figures (cue figure) is presented for a short time and the monkey judges whether the figure presented after a delay period is the match (target figure) or not.

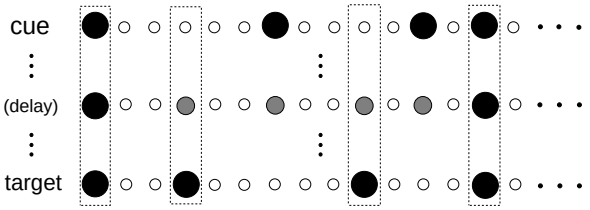

Fig. 1: Illustration of transition in the firing pattern during the recall process.

The result was basically the same as in the previous study by Miyashita and Chang [2], where the monkey memorized figures separately and sparse coding was found to be used for representing learned figures. However, two kinds of neurons exhibiting distinctive activities, called "pair-coding" and "pair-recall" neurons, were newly observed.

The former shows a selective response to both figures of a pair and exhibits a sustained activity during the delay period. The latter shows no response to the cue figure, but gradually increases its activity during the delay period and exhibits the maximum activity when the target figure is presented.

3. Interpretation and modeling

The above result can be interpreted as follows (see Fig. 1). Each figure is represented by a sparse firing pattern of a neuron group. This pattern does not depend on pictorial features of the figure, but paired cue and target figures are encoded into mutually similar patterns and their overlapping part corresponds to pair-coding neurons. During the delay pe-

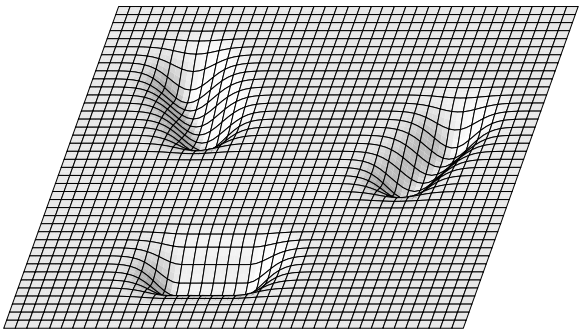

Fig. 2: Schematic energy landscape of a network storing paired associates.

riod, the firing pattern changes gradually from the cue-coding pattern to the target-coding pattern; in this process, some neurons act as pair-recall neurons.

To realize such a gradual shift of firing pattern stably, it is thought that not only cue-coding and target-coding states of the network but also the entire path connecting them should be smooth and attractive, or at the bottom of a "gutter of energy", as schematically depicted in Fig. 2. In this figure, the x-y surface represents the state space of the network and the z-axis represents the energy; three gutters corresponding to three pairs are drawn.

However, such a landscape can hardly be formed by conventional artificial neural networks that usually have a rippled energy landscape [3]. This problem was first solved by Morita [3] by improving network dynamics. Specifically, it has been shown that if we introduce neural elements with a nonmonotonic activation function, the energy landscape becomes smooth and trajectory attractors are easily formed.

Though nonmonotonic elements are biologically implausible, similar dynamics can be realized by combining a few normal monotonic elements as described below [4,5]. Moreover, this network accords with a broad distribution of sustained neuronal activities observed in the monkey inferotemporal cortex [2], which also is difficult to explain by the conventional models [4]. On these grounds, we use the following network for modeling the

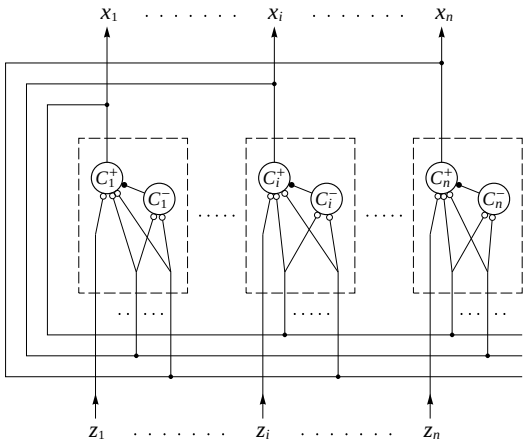

Fig. 3: Structure of the network.

pair-association memory.

4. Model

This model consists of pairs of excitatory and inhibitory cells, as shown in Fig. 3. A pair of cells surrounded by broken lines composes a unit, where the excitatory cell C_i^+ emits the output of the unit and the inhibitory cell C_i^- sends a strong inhibitory signal to C_i^+. Both cells receive recurrent inputs from other units. In mathematical terms,

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

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

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

where x_i and y_i are the outputs of C_i^+ and C_i^-, respectively, u_i is the potential, z_i is the external input, w_{ij}^+ and w_{ij}^- are synaptic weights from the j-th unit to C_i^+ and C_i^-, respectively, w_i^* represents the efficiency of the inhibitory synapse from C_i^- to C_i^+, and θ is a positive constant.

The activation function $f(u)$ of each cell is a monotonic sigmoid function increasing from 0 to 1. However, the input-output characteristics of the unit are nonmonotonic; that is, the output x increases with the total input v when v is small enough, but decreases when v becomes large and the inhibitory cell emits a large output.

4.1 Learning Algorithm

Learning is performed using a binary vector $\boldsymbol{r} = (r_1, \ldots, r_n)$ as a learning signal, which specifies a state to be memorized [5]. Specifically, we input $\boldsymbol{r}$ in the form $z_i = \lambda r_i$, where λ denotes input intensity, and modify synaptic weights according to

$$\tau' \frac{dw_{ij}^+}{dt} = -w_{ij}^+ + \alpha r_i x_j, \quad (4)$$

$$\tau' \frac{dw_{ij}^-}{dt} = -w_{ij}^- - \beta_1 r_i x_j + \beta_2 x_i x_j + \gamma. \quad (5)$$

Here α, 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 ($\tau' \gg \tau$). The coefficient α may be a positive constant, but the learning performance is better when α is a decreasing function of x_i; β_1 and β_2 are constants which satisfy $0 < \beta_1 < \beta_2$.

If the i-th unit receives a learning signal ($r_i = 1$) and its output x_i is small ($x_i < \beta_1/\beta_2$), then w_{ij}^+ is reinforced and w_{ij}^- is depressed, thus the output x_i increases. When x_i becomes large, however, w_{ij}^- is reinforced, thus x_i is restrained from increasing excessively. If $r_i = 0$, then only w_{ij}^- is reinforced, thus x_i decreases.

It should be noted that the term $\beta_2 x_i x_j$, or reinforcing w_{ij}^- according to the final output x_i of the unit, is indispensable for maintaining the nonmonotonic characteristics. Also, the term γ or lateral inhibition plays an important role in maintaining the total activity of units at a low level, allowing sparse coding.

Through this learning, the network energy is lowered around the state specified by $\boldsymbol{r}$, and thus the state becomes a point attractor if $\boldsymbol{r}$ is constant. If $\boldsymbol{r}$ changes successively at a slow pace, however, a gutter is engraved in the energy landscape along its track. In addition, a gentle flow in the same direction as the movement of $\boldsymbol{r}$ is generated at the bottom of the gutter. Consequently, the network state moves along the trajectory of $\boldsymbol{r}$ without external inputs only if a proper initial state is given [5].

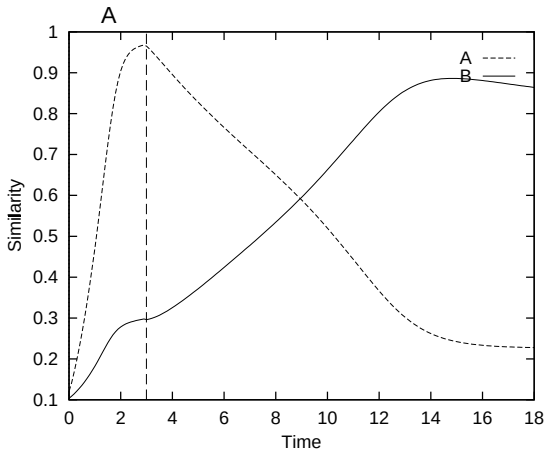

Fig. 4: Recall process when a cue A is given.

5. Computer Simulations

In this study, we examine whether the model can explain the above empirical data when an appropriate learning signal is externally given. For this purpose, we carried out computer simulations with 1000 units.

We prepared 20 pairs of patterns that are 1000-dimensional sparse vectors with 10% of elements being 1 and the rest 0. These patterns were randomly generated and have small correlations between different pairs, but paired patterns were selected such that they share 25% of the elements taking 1.

5.1 One-way recall

In the first simulation, one of a pair was always used as a cue-coding pattern and the other as a target-coding pattern. We then produced a spatiotemporal pattern which varied successively from the cue-coding to the target-coding pattern for each pair. Using this as the learning signal $\boldsymbol{r}$, we performed the above learning over 15 cycles, gradually decreasing the input intensity λ of $\boldsymbol{r}$.

Figure 4 shows a recall process after learning, which indicates the time course of change in the output vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ when a cue-coding pattern A is given during the cue period ($0 < t < 3\tau$) and no input is given thereafter. The ordinate of the graph is the similarity (direction cosine) between $\boldsymbol{x}$ and A or B (the target-coding pattern), and the ab-

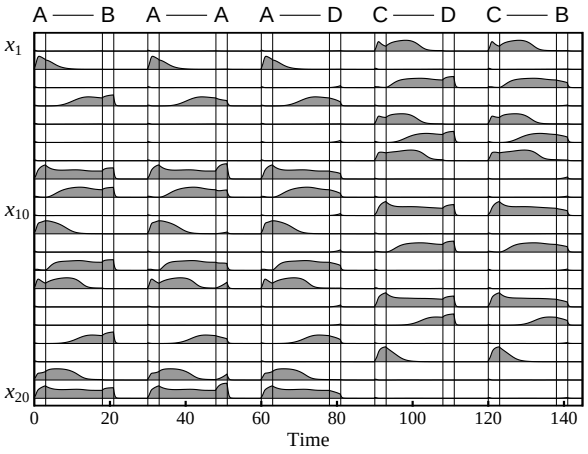

Fig. 5: Response of each unit to various inputs.

scissa is time scaled by the time constant τ.

We see that the output vector $\boldsymbol{x}$ varies gradually and B is recalled with a small error. In the same manner, it was confirmed that the correct target is recalled for all pairs.

Figure 5 shows the outputs of 20 units selected randomly out of the units showing some response, where five trials are performed every 30 τ. In each trial, cue A or C is given in the cue period and pattern A, B or D (target to C) is given in the choice period after a delay. The input level (average of z_i) is lowered in the choice periods and in the intervals between trials.

We see that many units exhibit a large output in the choice periods of the first and fourth trials in which the correct target is inputted, whereas the outputs are small when an unrelated pattern is inputted in the third and fifth trials; in the second trial, some units emit a large output when the cue A is re-inputted in the choice period, but they are few in number. Accordingly, we can clearly discriminate between correct and incorrect targets by the histogram of outputs, which indicates that the network is able to distinguish the correct target.

We also see that some units (e.g., nos. 8 and 10) respond to both cue and target and sustain a large output during the delay period, and others (e.g., nos. 4 and 6) show no response to the cue but strong response to the target

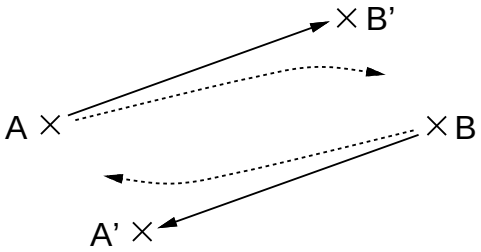

Fig. 6: Paths of the learning signal.

with a gradually increasing output in the delay period (note that the learning signal r_i took a binary value). These activities correspond well to those of the pair-coding and pair-recall neurons in the monkey brain.

5.2 Bidirectional recall

In the experiment by Sakai and Miyashita [1], monkeys were trained to recall either figure of a pair, in other words, cue and target figures were not fixed but interchangeable. However, we cannot realize such bidirectional recall simply by giving the same learning signal in reverse order. This is because the two paths of the learning signal $\boldsymbol{r}$ connecting paired patterns, for example from A to B and from B to A, overlap with each other and learning fails due to the interference between them.

We thus modify the paths of $\boldsymbol{r}$ as shown in Fig. 6; that is, we use spatiotemporal patterns varying from A to B' and from B to A' for association between A and B, where A' and B' are patterns with a moderate similarity to A and B, respectively. Using the modified $\boldsymbol{r}$, we performed a simulation with the same conditions as in the previous one (1000 units, 20 pairs) but learning was performed 20 times for each direction. Patterns A' and B' were selected randomly out of the patterns having a similarity of 0.5 to A and B, respectively.

Behavior of the model for cues is shown in Fig. 7 and Fig. 8. We see that the output vector $\boldsymbol{x}$ approaches the correct target whichever pattern is given as a cue. Although the target is not completely recalled, the units coding the target exhibit a large response in the choice period because $\boldsymbol{x}$ becomes close enough

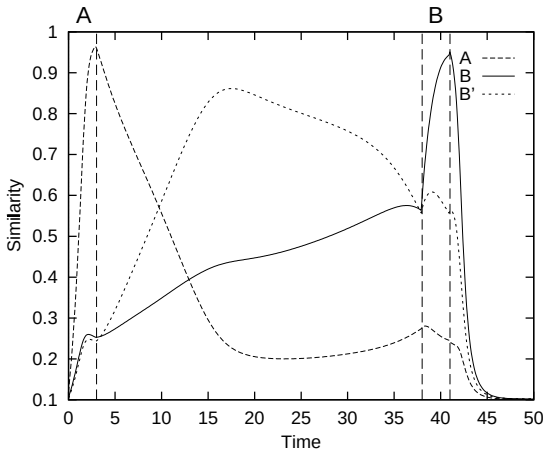

Fig. 7: Recall from A to B.

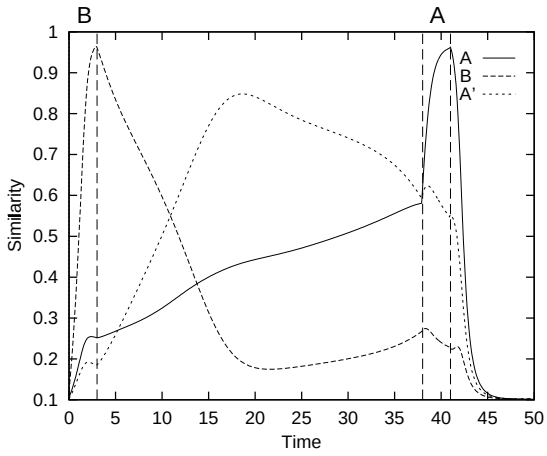

Fig. 8: Recall from B to A.

to the target pattern.

It should be noted that the similarity of $\boldsymbol{x}$ to B' (A') first increases to 0.85 but decreases thereafter, while that to B (A) increases consistently. This indicates that $\boldsymbol{x}$ follows the trajectory shown by the broken line in Fig. 6, because of attractive force between the two paths.

We confirmed that bidirectional recall is possible for all the pairs, and that some units exhibit similar activities to the pair-coding and pair-recall neurons as seen in the previous simulation.

6. Concluding remarks

We have discussed the mechanism of pair-association memory and demonstrated that a neural network model consisting of pairs of excitatory and inhibitory cells can explain the neuronal activities observed in the monkey inferotemporal cortex. This model uses a simple learning rule and is biologically plausible. We consider, therefore, that the model captures a fundamental principle of the pair-association memory in the primate neocortex.

However, we assumed in this paper that an appropriate learning signal is given externally to the model. Thus, if our view is correct, such a learning signal must be generated somewhere in the brain for the inferotemporal cortex. If so, where and how is it generated?

We believe that the learning signal is generated through cortico-hippocampal interactions. This idea is supported by recent findings of Miyashita *et al.* [6] that the above pair-association neurons are not found in monkeys with lesions of the entorhinal and perirhinal areas, which mediate mutual connections between the temporal cortex and the hippocampus. A revised model considering the hippocampal system (including the parahippocampal region) will be presented in the future.

References

[1] Sakai, K. and Miyashita, Y.: Neural organization for the long-term memory or paired association, *Nature*, **354**, 152–155 (1991).

[2] Miyashita, Y. and Chang, H. S.: Neuronal correlate of pictorial short-term memory in the primate temporal cortex, *Nature*, **331**, 68–70 (1988).

[3] Morita, M.: Memory and learning of sequential patterns by nonmonotone neural networks, *Neural Networks*, **9**, 1477–1489 (1996).

[4] Morita, M.: A neural network model of the dynamics of a short-term memory system in the temporal cortex, *Systems and Computers in Japan*, **23**, 4, 14–24 (1992).

[5] Morita, M.: Computational study on the neural mechanism of sequential pattern memory, *Cognitive Brain Research*, **5**, 137–146 (1996).

[6] Miyashita, Y., Okuno, H., Tokuyama, W., Ihara, T., and Nakajima, K.: Feedback signal from medial temporal lobe mediates visual associative mnemonic codes of inferotemporal neurons, *Cognitive Brain Research*, **5**, 81–86 (1996).