ELSEVIER

Cognitive Brain Research 13 (2002) 169–178

COGNITIVE BRAIN RESEARCH

www.elsevier.com/locate/bres

Research report

Computational modeling of pair-association memory in inferior temporal cortex

Masahiko Morita[a,*], Atsuo Suemitsu[b]

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

Accepted 5 October 2001

Abstract

Distinctive neuronal activities related to visual stimulus–stimulus association have been found in the inferior temporal (IT) cortex of monkeys. They provide an important clue to elucidating the memory mechanisms of the brain, but do not accord with existing neural network models. In the present paper, we clarify the computational principle required for reproducing the empirical data and construct a biologically feasible model that learns and performs a delayed pair-association task. This model is composed of two neural networks, association network N_1 and trainer network N_2, and pair-association memories are formed by their interactions. Specifically, N_2 receives the output of N_1 in addition to an external input, and sends a learning signal back to N_1; this signal works as a guide for shifts in output pattern or state transitions of N_1, and memory traces are engraved along its path, so that a trajectory attractor connecting from the cue-coding to the target-coding state is formed in N_1. Computer simulation shows that the model not only distinguishes the target in the task, but also explains the activity of the IT neurons very well. It is reasonable to presume that N_1 and N_2 correspond to area TE and the rhinal cortex, respectively; based on this theory, we explain some physiological findings on learning and memory, and also make several predictions.

Theme: Neural basis of behavior

Topic: Learning and memory: systems and functions

Keywords: Inferotemporal cortex; Pair-recall neuron; Perirhinal cortex; Computational theory; Trajectory attractor; Feedforward inhibition

1. Introduction

It is widely accepted that the inferior temporal cortex (IT) is deeply involved in visual memory, but how visual information is structured and transformed into long-term memory, and how such memory is retrieved and used for purposive behavior are not clear. One of the most important clues to these questions may be the finding by Sakai and Miyashita [16]. They recorded IT neurons of the monkey and found novel activities related to visual stimulus–stimulus associations, offering a crucial hint on how paired associates are encoded and recalled in IT. However, the neural mechanism underlying these memory processes is not yet known, nor does there exist a computational model that adequately explains the empirical data. In the present study, we construct a neural network model of pair-association (PA) memory by a computational approach to this problem, and comparing the model with physiological findings, we examine the computational principle of memory in the temporal lobe.

1.1. Interpretation of PA-related neurons

In the experiment by Sakai and Miyashita [16], monkeys were trained on a delayed pair-association (DPA) task using 12 pairs of computer-generated pictures. In each trial of the task, one of the pictures is presented as a cue, and the monkey must judge whether a test picture presented after a delay interval is the paired associate (target) of the cue or not. The results of this experiment may be summarized as follows.

*Corresponding author. Tel.: +81-298-53-5321; fax: +81-298-53-6554.

E-mail address: mor@bcl.esys.tsukuba.ac.jp (M. Morita).

0926-6410/02/$ – see front matter
PII: S0926-6410(01)00109-4

First, most of the stimulus-selective neurons respond to a few pictures, and these pictures apparently have no obvious pictorial features in common, which is the same as in the previous study using a delayed match-to-sample (DMS) task [7,8]. Second and more important, two kinds of characteristic neurons, termed 'pair-coding' and 'pair-recall' neurons, were observed. The former shows a selective response to both pictures of a pair, usually exhibiting a sustained activity during the delay period. The latter shows no response to the cue, but gradually increases activity during the delay period, exhibiting the maximum activity when the target is presented. It should be noted that the same neuron can exhibit both types of activity for different picture pairs.

When we consider this result from a viewpoint of systems, the following interpretation may be the most plausible (see Fig. 1).

1. Each picture is represented by a firing pattern of a neuron group in IT. This pattern ('code' of the picture) is sparse in that active neurons are small in number, and individual neurons do not encode a particular feature of the picture. Such a manner of representation is called sparse representation.
2. The codes for paired pictures have some similarity to each other, and the overlapping part corresponds to pair-coding neurons.
3. After cue presentation, the firing pattern changes gradually from the cue-coding to the target-coding pattern. In this process, some neurons act as pair-recall neurons.

This interpretation, particularly the gradual shift of the firing pattern during the delay period, implies that the neurons compose a dynamical system as schematically depicted in Fig. 2. That is, not only cue-coding and target-coding states of the system but also the entire path connecting them should be attractive, or at the bottom of a 'gutter of energy', in order that the system may stably maintain the firing pattern during the course of state transition. In addition, the bottom of the gutter should be smooth with some 'flow' toward the target-coding state in order that a continuous state transition may be achieved without stopping in the middle. Such an energy gutter is called a trajectory attractor.

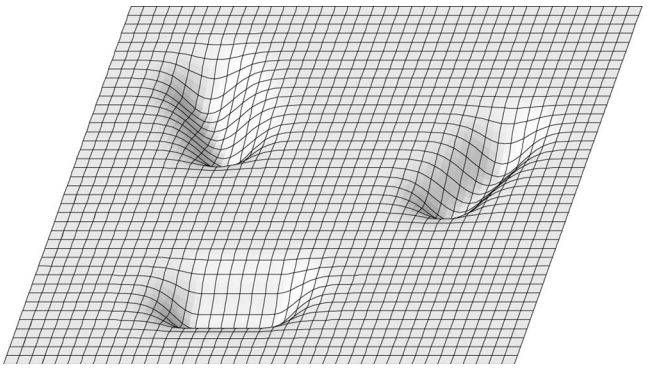

Fig. 2. Schematic energy landscape of a network storing pair-association memories. The surface represents the state space of the system which is actually of very large dimensions, where every point on the surface corresponds to a firing pattern and neighboring points correspond to very similar patterns. The height represents potential energy indicating stability of each state, and the current state (firing pattern) of the system changes toward a state with lower energy if no external input is fed; thus a state with lower energy than the neighboring states is stable, called an attractor. Three string-shaped attractors are drawn, each of which corresponds to a memory trace associating a cue with the target.

Thus far, many neural network models with point attractors, namely isolated energy holes, have been presented. In particular, the studies by Amit and his colleagues (e.g. Ref. [1]) are in the same line as the present study in modeling neuronal activities in IT using attractor networks. These studies, however, deal mostly with DMS but not DPA tasks. It seems that point attractor networks

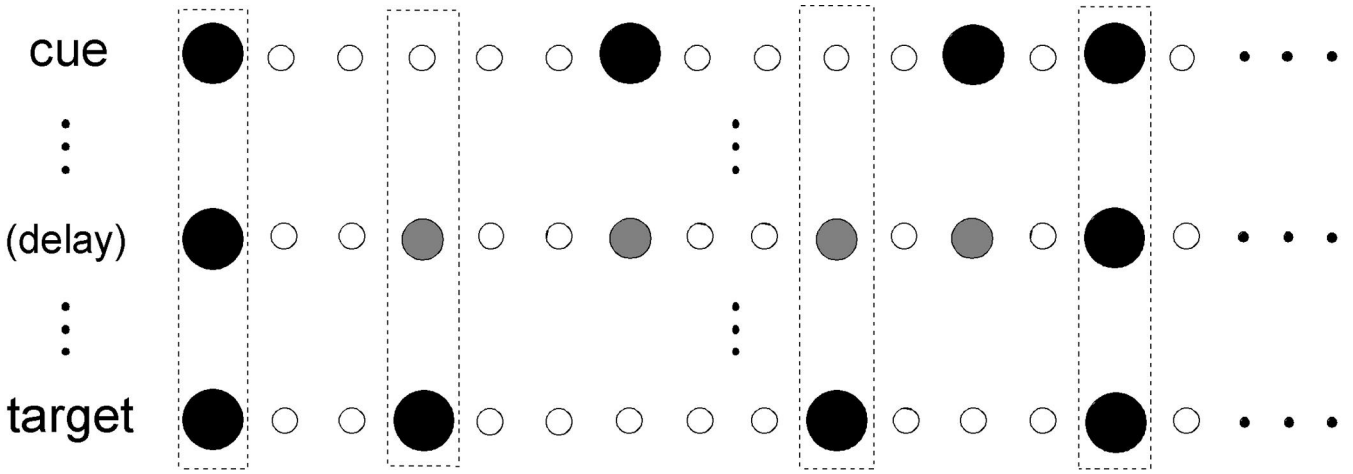

Fig. 1. Illustration of a presumed recall process in IT. The firing pattern of a neuron group representing the cue gradually shifts during the delay period into a different pattern representing the target.

cannot give a plausible account of the activity of pair-recall neurons, and we believe that trajectory attractors have to be introduced.

1.2. Constructing a trajectory attractor network

If the above view is correct, it becomes a critical issue how a dynamical system as shown in Fig. 2 can be realized in a feasible way, since conventional attractor neural networks, like the Hopfield model, generally have a rippled configuration of energy. In regard to this problem, a network consisting of elements with a nonmonotonic activation function shown in Fig. 3 is known as a nonmonotonic model [9], in which trajectory attractors can be formed using a Hebb-like learning rule [10]. Although this model is simple and performs well, nonmonotonic elements are not biologically plausible and are unsuitable for modeling the neuronal activity in IT. Instead, we adopt the feedforward-inhibition network model [11] shown in Fig. 4.

This model consists of interconnected units composed of a pair of excitatory and inhibitory cells. In the ith unit, both cells receive recurrent inputs from the other units they have in common, but only excitatory cell C_i^+ emits an outward signal; the output of inhibitory cell C_i^- is sent to C_i^+ through a strong inhibitory connection. In mathematical terms,

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

$$\tau \frac{\mathrm{d}u_i}{\mathrm{d}t} = -u_i + \sum_{j=1}^{n} w_{ij}^+ x_j - w_i^* y_i + z_i, \tag{2}$$

$$x_i = f(u_i), \tag{3}$$

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 the synaptic weights from the jth unit to C_i^+ and C_i^-,

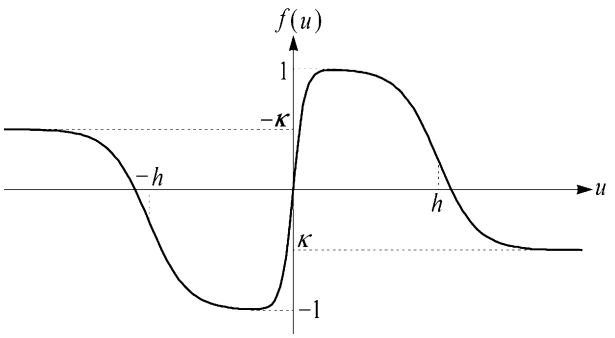

Fig. 3. Activation function for a nonmonotonic network model. This function designates the input–output characteristics of each element. By use of such a nonmonotonic function instead of the conventional sigmoid function, a recurrent neural network is markedly improved in many respects, one of which is that string-type attractors can be easily formed. The detailed shape of the function $f(u)$ is not very critical but it is essential that $f(u)$ decreases with u when $|u|$ is large.

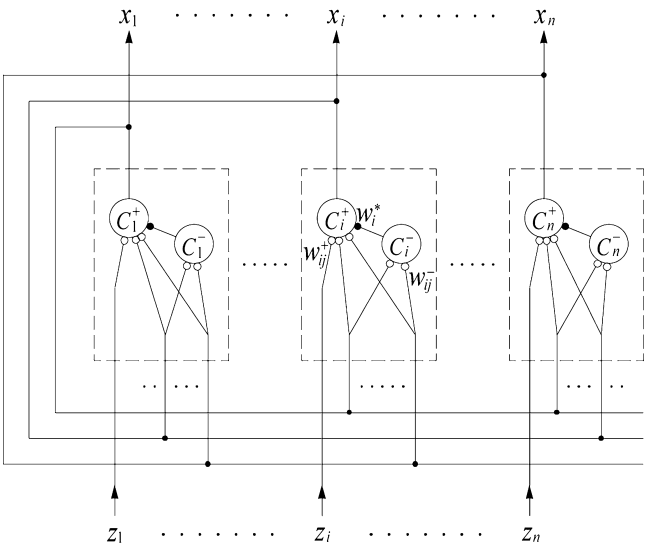

Fig. 4. Structure of the feedforward-inhibition network. A pair of cells surrounded by broken lines composes a unit, and interconnected units compose the network, in which trajectory attractors can be formed as in the nonmonotonic model but without using nonmonotonic elements.

respectively, w_i^* is the synaptic weight from C_i^- to C_i^+, and τ and θ are positive constants representing a time constant and a threshold, respectively.

The activation function $f(u)$ of each cell is a monotonic sigmoid function increasing from 0 to 1 given by

$$f(u) = \frac{1}{1 + \mathrm{e}^{-cu}}, \tag{4}$$

c being a positive constant. However, the input–output characteristics of the unit are nonmonotonic, since the output x_i begins decreasing with the total input when the inhibitory cell is activated to emit a strong inhibitory signal. It should be noted that this is the simplest composition of monotonic cells that realizes the same dynamical properties as the nonmonotonic model. Moreover, this model is very suitable for memory of sparse patterns because the feedforward inhibition has a function of preserving the total activity of the network at a constant low level [11].

Learning of this network to form a trajectory attractor is performed using a learning signal vector $\boldsymbol{r} = (r_1, \ldots, r_n)$, which changes gradually leading the state transition of the network and engraving a gutter of energy. Specifically, while the network is running according to Eqs. (1–3), each unit receives r_i in the form $z_i = \lambda r_i$, where λ denotes input intensity, modifying the synaptic weights according to

$$\tau' \frac{\mathrm{d}w_{ij}^+}{\mathrm{d}t} = -w_{ij}^+ + \alpha r_i x_j, \tag{5}$$

$$\tau' \frac{\mathrm{d}w_{ij}^-}{\mathrm{d}t} = -w_{ij}^- - \beta_1 r_i x_j + \beta_2 x_i x_j + \gamma. \tag{6}$$

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 constant, but the learning performance is better when α is a decreasing function of x_i; β_1 and β_2 are positive constants.

The process of learning is schematically shown in Fig. 5. Intuitively, the above synaptic modification lowers the energy around the state specified by $\boldsymbol{r}$; thus a point attractor is formed if $\boldsymbol{r}$ is fixed. When $\boldsymbol{r}$ moves successively at a slow pace, however, the network state $\boldsymbol{x}$ follows slightly behind it, and a gutter is engraved along the track. In addition, the small gap between $\boldsymbol{r}$ and $\boldsymbol{x}$ produces a weak flow from $\boldsymbol{x}$ toward $\boldsymbol{r}$, namely in the same direction as the movement of $\boldsymbol{r}$, at the bottom of the gutter. By repeating this process several times, a trajectory attractor is formed along the trajectory of $\boldsymbol{r}$.

1.3. Problems in modeling

As described above, a dynamical system as expressed by Fig. 2 can be modeled with a neural network. For modeling the PA memory in IT, however, the following problems must be solved.

1. In the actual task, the cue of a trial is randomly chosen between the two pictures of a pair, so that the monkey has to recall either picture from the other. In the model, however, the state transition along a trajectory attractor proceeds only in a fixed direction. Even if we overlap an additional training in the reverse direction, it results in failure because of interference.
2. In training the monkey, no stimulus changing gradually from the cue to the target is given, but only cue and target pictures are presented. The model can merely store separate memories and cannot form a trajectory attractor if the learning signal changes very rapidly or jumps abruptly from one pattern to a quite different one.
3. The monkey is required to discriminate between target and nontarget pictures. Accordingly, the model should possess a natural mechanism of target recognition. Simply comparing the input and recalled patterns for every component is not biologically plausible.

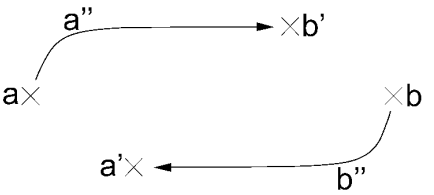

Fig. 6. Paths of the learning signal. The n-dimensional vector space is expressed two-dimensionally, where a and b represent code vectors of paired associates.

Among these, problem 1 can be solved by entirely separating the trajectories for two recall directions. This is achieved by using a learning signal that changes from a to b′ and from b to a′ as shown in Fig. 6, where a and b represent picture-coding patterns and a′ and b′ are moderately different patterns from a and b, respectively. It follows, however, that b is not exactly recalled even if a is given to the network.

Nevertheless, problem 3 means conversely that the target does not require exact recall as long as it is

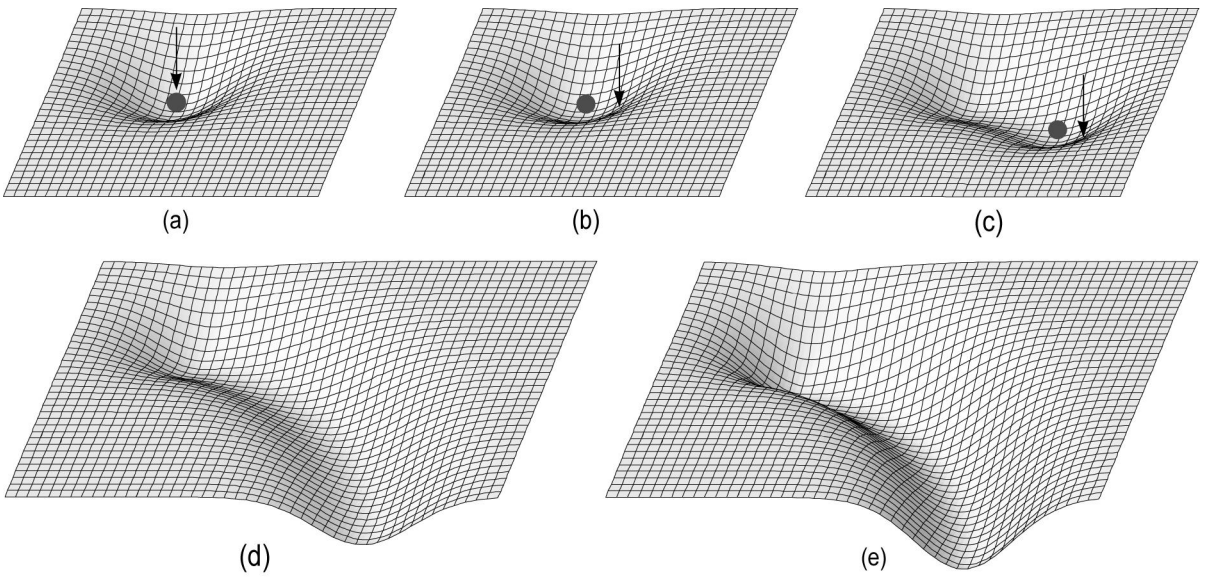

Fig. 5. Illustration of the learning process for forming a trajectory attractor. The change in the 'energy landscape' is depicted in the order (a) to (e). The solid circle represents the current state $\boldsymbol{x}$ of the network and the arrow represents the current learning signal $\boldsymbol{r}$. As $\boldsymbol{r}$ gradually moves, $\boldsymbol{x}$ follows behind and a string-shaped attractor is formed along the track.

distinguishable from nontargets. In fact, as described later, the feedforward-inhibition network can make a correct discrimination without exact recall of the target in a natural manner [17]. Consequently, the above problems are condensed into the problem of how to internally generate a desirable learning signal as shown in Fig. 6 when cue and target patterns are given separately with an interval.

The easiest solution to this problem may be introducing another network that transforms the input patterns into the learning signal **r**. We adopt this method and refer to the additional network as a trainer network denoted by N_2, whereas the feedforward inhibition network in which PA memories are formed is termed an association network denoted by N_1.

Actually, however, it is very difficult to realize such transformation by a single network, in that **r** should vary slowly along a long trajectory. Long state transition in an independent network is generally incompatible with slow continuous transition, except for trajectory attractor networks which require a learning signal again. A single transformation network is also undesirable in that the output pattern **r** is determined irrespective of the state **x** of N_1, although **r** has to lead **x**.

2. The model

The above discussion suggests that difficulties in generating the learning signal can be resolved if networks N_1 and N_2 are interactive. Based on this idea, we constructed the following model, with N_2 being simplified as much as possible.

2.1. Structure

Fig. 7 shows the composition of the model, where trainer network N_2 not only sends **r** to association network N_1 but also receives **x** from N_1. Although N_1 as well as N_2 should receive the input pattern **s**, the direct input to N_1 is not used in the present model.

The structure of trainer network N_2 is shown in Fig. 8. This network consists of n cells having one-to-one correspondence to the n units of N_1. The ith cell C_i receives the input pattern $\boldsymbol{s} = (s_1, \ldots, s_m)$ through synaptic weights p_{ij} and emits r_i to the ith unit of N_1. The synaptic weight p_{ij} individually takes a random value so that N_2 works as a random transformation network. Cell C_i also receives a feedback signal x_j from every unit of N_1 through a random synaptic weight q_{ij}.

This network is a kind of competitive system as well, since C_i has self-excitatory and lateral inhibitory connections. This permits only a few cells to emit a large output whereas the other outputs are almost zero, which means **r** is a sparse pattern.

In mathematical terms,

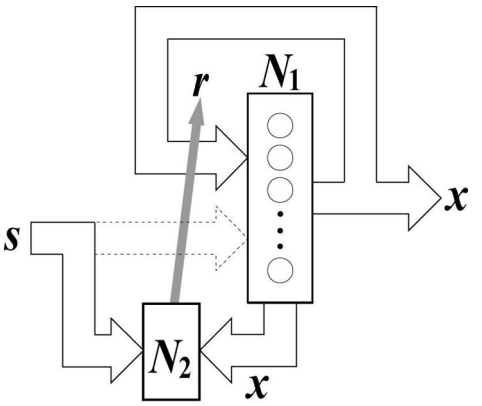

Fig. 7. Block diagram of the model. Association network N_1 receives the output **r** of trainer network N_2, and N_2 receives the output **x** of N_1. External input **s** is fed only into N_2 and the direct input path to N_1 (broken lines) is omitted for simplicity. For the specific structure of N_1 and N_2, see Figs. 4 and 8, respectively.

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

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

where v_i denotes the potential of C_i, ρ and σ are positive constants representing the efficiency of lateral inhibition and self-excitation, respectively, and η is an offset.

2.2. Behavior

To understand the behavior of the model in learning, relation and interaction between the two networks are important. Since N_1 receives **r** in the form $z_i = \lambda r_i$, its output vector x is generally similar (except in terms of magnitude) to vector r. When **r** varies, however, x follows **r** at some interval; x also differs from **r** in that it shifts

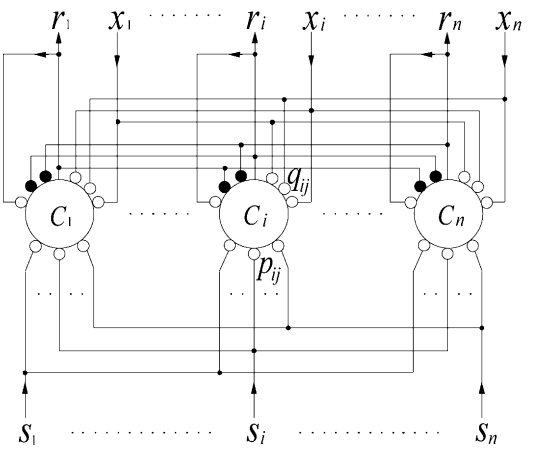

Fig. 8. Structure of the trainer network N_2. Each cell C_i has lateral inhibition and self-excitation, and outputs r_i by receiving external input **s** and the output **x** of N_1 through random synaptic connections.

gradually even if $\boldsymbol{r}$ changes very rapidly or discontinuously. On the other hand, the random connections from N_1 to N_2 have the action of driving $\boldsymbol{r}$ in various directions depending on $\boldsymbol{x}$, but do not cause an abrupt change in $\boldsymbol{r}$ since N_2 is a competitive network in which winner cells tend to maintain their activity.

Based on this argument, let us assume that we input a cue pattern A to the model in the rest state (in which all cells are inactive) and the target pattern B after a delay. First, when A is fed to N_2, it is transformed into a sparse pattern a and sent to N_1, and shortly thereafter the state of N_1 becomes a. Although $\boldsymbol{x}$ is fed back to N_2, $\boldsymbol{r}$ is almost constant while $\boldsymbol{s}$ = A. When the input of A ends ($\boldsymbol{s} = 0$), the feedback signal from N_1 becomes relatively dominant and $\boldsymbol{r}$ is moved from a to a somewhat different pattern a″ (see Fig. 6). If N_2 is in the rest state, $\boldsymbol{r}$ becomes b immediately after the input of B; however, because B is fed while N_2 is emitting a″, $\boldsymbol{r}$ moves gradually to a pattern b′ which lies in between a″ and b.

In the same way, by feeding B and A in this order, a learning signal changing gradually from b via b″ to a′ is generated. It should be noted that the feedback connections from N_1 to N_2 are essential not only for regulating the moving rate of $\boldsymbol{r}$ but also for separating the two paths of $\boldsymbol{r}$ in Fig. 6.

In parallel with this process, learning of N_1 is performed using $\boldsymbol{r}$, as described previously, so that trajectory attractors along these paths are formed. As a result, when we input A as a cue and a is sent to N_1 through N_2, $\boldsymbol{x}$ shifts to b′ during the delay period. If we then input B, $\boldsymbol{x}$ quickly changes to b, and thus N_1 is thought to show a strong response.

Incidentally, when the model performs the task after learning, input patterns A and B need to be transformed into their codes a and b as they are in learning. Although we use N_2 for this transformation for simplicity of the model, we may directly input $\boldsymbol{s}$ to N_1 (see Fig. 7) and train the input synapses to transform $\boldsymbol{s}$ into its code; then it is possible for N_1 to perform the task without N_2.

3. Computer simulation

Computer simulation was carried out using a network with a size of $n = 1000$. First, we randomly generated 20 pairs of patterns that are 1000-dimensional ($m = 1000$) sparse vectors with 10% of elements being 1 and the rest 0. We then input a pair of patterns in some order applying the above learning procedure, and after resetting the model to the rest state, we input them in the reverse order and training was conducted. After resetting the model again, the other pairs were fed in the same way, which composes one cycle of training.

Parameters were adjusted by several trials to set

$\theta = 3,\ w_i^* = 10,\ \tau' = 50000\tau,\ c = 10,\ \lambda = 0.3,$
$\beta_1 = 25,\ \beta_2 = 50,\ \gamma = 0.05,\ \rho = 0.016,\ \sigma = 0.8.$

We set $\alpha = \phi(x_i)$, where the function $\phi(x) = 50(0.5 - x)$ for $x \leq 0.5$ and 0 for $x > 0.5$, and η to be normally 0 but 0.75 during the delay period in learning. Synaptic weights p_{ij} and q_{ij} were set to random numbers with mean 5.5×10^{-3} and variance 3.7×10^{-4} and with mean 7.2×10^{-3} and variance 6.4×10^{-4}, respectively.

After completing 20 cycles of training, we tested the model by repeating a trial in which we gave a cue pattern to the model, and input a test (target or nontarget) pattern after a delay, varying combination of the cue and test; an inhibitory signal sufficiently strong to reset the network N_1 is fed during intertrial intervals. Response of the model is shown in Fig. 9, where the time course of the outputs of 20 units in N_1 is plotted. These units were randomly selected from among the units that encode some of the patterns A to D, but those displaying similar behavior were omitted. The second, fifth, tenth and twelfth trials are match trials in which the test pattern is the target, and the others are nonmatch trials.

Comparing the first four trials in which A is the cue, we see that most of the active units further increase their output when B is fed in the test period, but are depressed for C or D; even if A is fed again as a test pattern, the response of the units is not so strong as that to B. In contrast, the response is strongest to A when B was given as a cue. Similarly, the target elicits a stronger response than nontargets in the case that C or D is the cue.

In relation to this, such enhancement of response to the match stimulus is observed also in IT and thought to be involved in mechanisms of recognition [5,6]. However, the response modulation in IT is not fully explained by the model, since many IT neurons show a suppressed response to the match stimulus, and moreover, the enhancement and suppression effects are maintained even after intervening stimuli.

Fig. 10 shows histograms of the outputs of all units to test input, where (a) and (b) are those in the second and third trials in Fig. 9 and are typical cases of match and nonmatch trials, respectively. Although their averages are nearly equal due to the total activity control property of the feedforward inhibition network, the two distributions are obviously different. In fact, the number of units with an output of more than 0.5 is about three times larger in (a) than in (b), and we confirmed that by this difference, match and nonmatch responses were distinguishable for all combinations of cue and test patterns. This indicates that the model is able to recognize the target.

In addition, we can see that the units exhibit similar activities to IT neurons. For example, unit no. 20 responds to both A and B and sustains a substantial output during the delay period, which corresponds well to the pair-coding neuron. There also exist many units that exhibit no response to the cue but strong response to the target with a gradually increasing output in the delay period, in the same way as the pair-recall neuron does. These units decrease their activity during the delay period if cue and target

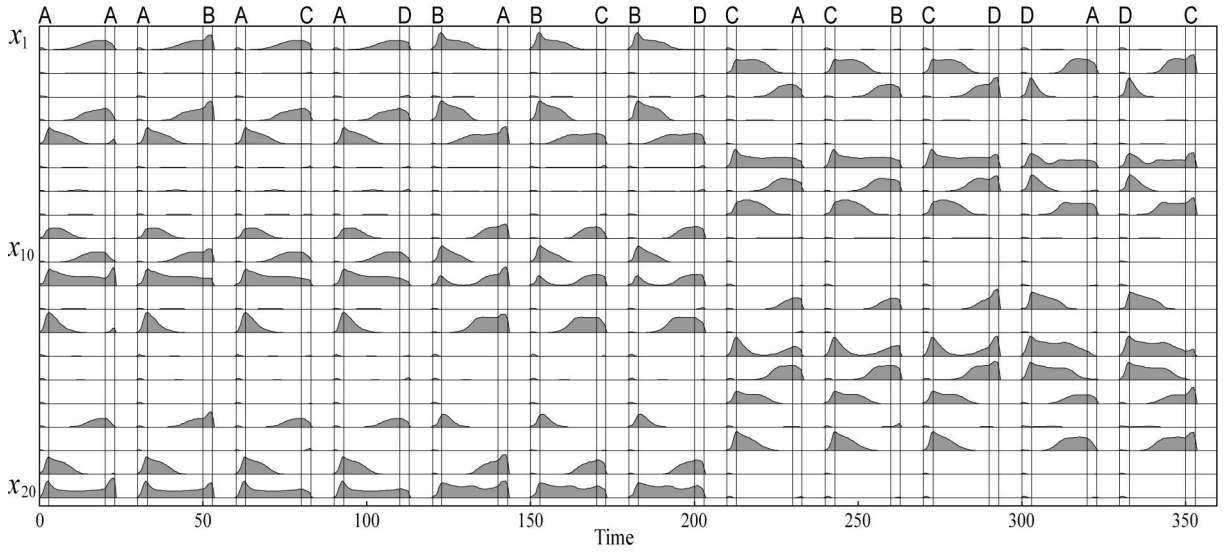

Fig. 9. Behavior of the model after learning. Responses of individual units in N_1 to various cue and test patterns are vertically arranged. Twelve trials are shown, each of which consists of cue, delay and test periods. The abscissa is time scaled by the time constant τ in Eq. (2).

patterns are interchanged, which also applies to the pair-recall neuron.

4. Discussion

As described above, this model can not only perform the DPA task, but also reproduce the activity of IT neurons well. Moreover, this model is constructed on the basis of computational requirements, its working principle is biologically feasible even in a huge-scale network, and furthermore, no other computational model, at least currently, can satisfactorily explain the above empirical data. We therefore believe that the same principle underlies the neural mechanism of PA memory in the temporal lobe.

If our view is correct, the structure of the brain must be reflected in that of the model to some extent, although it was not directly referred to in constructing the model. Then, what correspondence can be found between the two? To answer this question, we should note the following.

First, since the interaction between N_1 and N_2 is important in the model, corresponding brain areas should have strong neural connections with each other.

Second, trainer network N_2 is necessary for forming

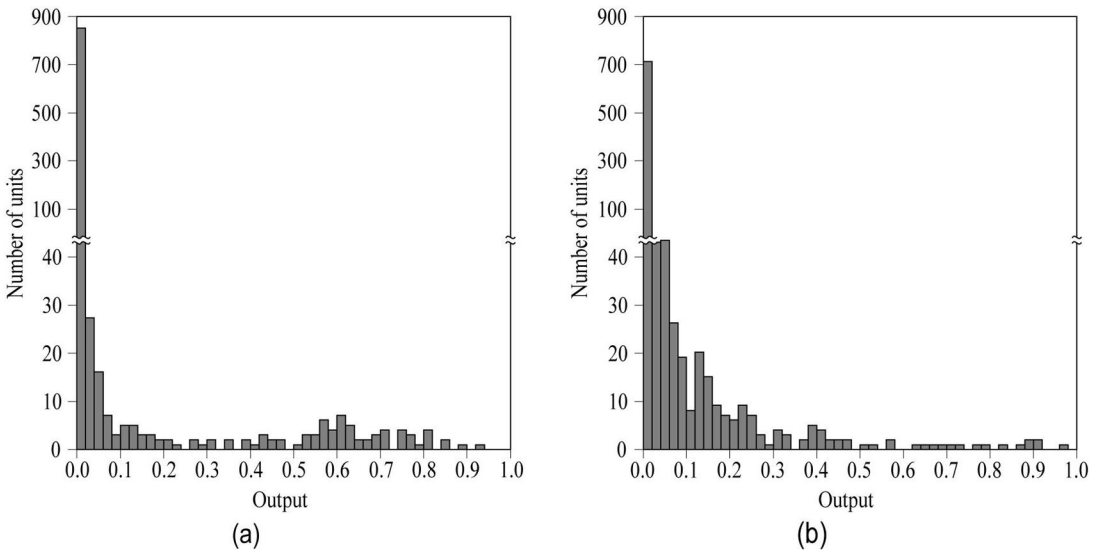

Fig. 10. Responses of the model to (a) target and (b) nontarget patterns. Distribution of the outputs of all units at the end of the test period is shown as a histogram. The number of units with an output of more than 0.5 is 56 in (a) but 19 in (b).

trajectory attractors or PA memories in association network N_1, but not for forming point attractors or separate memories; it is also not essential for recognition of the target if training of N_1 including the direct input path has been completed.

Third, Murray et al. [12] demonstrated that monkeys with lesions of the rhinal cortex (perirhinal and entorhinal cortices) cannot perform new visual stimulus–stimulus learning at all, although recognition of visual stimuli is intact. Interestingly, the monkey was able to re-learn the association that had been learned before the lesion.

Fourth, Higuchi and Miyashita [4] trained monkeys on the above DPA task with section of the anterior commissure plus ablation of rhinal cortex in one hemisphere, and recorded neuronal activity in area TE of both hemispheres. They reported that although the monkeys could correctly perform the task, PA-related neurons were observed only in the intact hemisphere; TE neurons in the lesioned hemisphere exhibited stimulus selectivity but not pair-coding or pair-recall activity.

Considering these facts, together with the fact that the rhinal cortex (especially the perirhinal cortex) is anatomically adjacent and strongly interconnected to area TE, it is reasonable to presume that networks N_1 and N_2 of the model correspond to area TE and the rhinal cortex of the temporal lobe, respectively.

In relation to this, according to Eichenbaum et al. [2], the rhinal cortex should be included in the hippocampal system. The hippocampus itself, however, does not correspond to N_2 of the present model because visual PA learning in monkeys is not disrupted by hippocampal removal alone [12]. Nevertheless, as pointed out by Eichenbaum et al. [2], it is likely that both the hippocampus and rhinal cortex participate in PA learning in different ways. Thus it may be possible to model a part of hippocampal function by examining the requirements for N_2 in performing a more difficult task affected by hippocampal damage.

Finally, we list in the following some other phenomena explained or predicted by the model.

(1) Assume that stimuli A and B and stimuli C and D are paired associates, respectively, and that A and C are highly similar. In this case, errors are known to increase markedly in trials such that either A or C is the target and the other is used as a test stimulus, whereas the error rate does not significantly rise when A or C is used as a cue [15]. This phenomenon is explained by the model as follows. If A and C are similar patterns, their codes a and c in N_1 are also close. Even in this case, trajectory attractors as shown in Fig. 6 can be formed without particular problems. Then, if N_1 receives a, it recalls b′ as in normal cases so that no error arises from the similarity between a and c. If B is given as a cue, however, N_1 recalls a′ that is near not only to a but also to c so that the model exhibits a considerably strong response to the test input of C. This will cause an error in recognition of the target, since a is at a distance from a′ and the response elicited by the test input of A is not overwhelmingly stronger. If the target-coding pattern were recalled exactly, such asymmetry between similar cues and similar targets would be greatly reduced. Thus, the model is consistent with the behavioral data, implying that the code of a cue stimulus and that recalled from the paired associate are not identical in the brain as well.

(2) In our model, interactions between N_1 and N_2 are necessary mainly for interpolating the cue-coding and target-coding states to form a continuous trajectory attractor. Such interpolation is easier if the two states are close. Accordingly, giving a subtarget midway between the cue and target will facilitate learning. In fact, in the experiment by Murray et al. [12], they first trained monkeys using a compound stimulus consisting of the target superimposed on the cue, so that the PA learning was greatly promoted. If we extend this, by presenting sequential stimuli varying gradually from the cue to the target, it may be possible to train a monkey with rhinal cortex lesions to learn the association to some extent. However, even if it is possible, bidirectional (A→B and B→A) learning will be more difficult for lesioned monkeys than one-way (A→B only) learning, since the interactions between N_1 and N_2 also play a role of reducing interference between associations in the opposite directions.

(3) As seen in Fig. 9, the units of the model with increasing output during the delay period become active at various times. This is a direct reflection of a gradual shift of the state of N_1, since if it jumps, many units should change their output synchronously. Accordingly, the model predicts that the onset time of the pair-recall activity in area TE will be highly diverse.

(4) Comparing the second and fifth trials in Fig. 9, we can see that the course of the outputs for cue B is not entirely the time reversal of that for cue A. This is because the paths of the state transition of N_1 in the two cases are different as previously described. It follows, therefore, that many pair-recall neurons in IT should exhibit such asymmetrical activity.

(5) Although unit no. 14 in Fig. 9, for example, encodes both of the paired patterns C and D, its output decreases during the delay period when C is given as a cue. This reflects a ‘curved’ trajectory of the state of N_1 as shown in Fig. 6 (during a ‘straight’ transition, every unit shows a monotonically varying or nearly constant output). This also leads to a prediction that some of the pair-coding neurons in IT do not exhibit a sustained activity during the delay period.

(6) When cue B is given to the model, not all units encoding A increase their output during the delay period; also, some of the active units are depressed by the test input of A. These phenomena arise from the above discrepancy between a and a′, implying that similar neuronal activities will be seen in IT.

(7) Inhibitory cells of the model have very different

properties from the units or excitatory cells, in that they usually do not show a strong response to a particular input pattern, and that their output is usually at a low level, not varying very much during the delay period. These properties may give the impression that the inhibitory cells are not important, but they have an essential role as previously described. Accordingly, at least some part of the IT neurons exhibiting little stimulus selectivity can be explained by the model.

Incidentally, Naya et al. [14] quite recently reported on the time course of pair-recall activity of neurons recorded from area TE and also from area 36, a part of the perirhinal cortex. They calculated a pair-recall index (PRI) and examined how PRI varies during the delay period. According to their data, PRI clearly increases but does not exceed 0.5. This implies that the target response and the recall activity are similar but substantially different, supporting the prediction in (6).

More importantly, their data showed that the onset time of pair-recall activity of TE neurons is distributed nearly uniformly over a rather wide range, which verifies the prediction in (3). Naya et al. also demonstrated that pair-recall activity in area TE is distinctly preceded by that in area 36, although the simple cue response is slightly faster in area TE. This remarkable finding also supports our hypothesis that the learning signal necessary for forming trajectory attractors is sent from the rhinal cortex to area TE, since it leads, and thus should precede, the state transition of the association network (see Fig. 5).

However, the present model has some limitations. First, an increase in the ratio of pair-coding neurons during the process of learning, which seems crucial for the structuring process of memory, is not fully explained. To solve this problem, introducing some plasticity into N_2 of the model will be necessary. Conversely, N_2 is insufficient as a model of the rhinal cortex or hippocampal system, since this area has a complex structure with rich plasticity.

Second, the model does not deal with the function of the prefrontal cortex (PFC), which is known to participate in PA tasks. For example, Rainer et al. [15] found neurons in PFC that exhibit activity reflecting the anticipated target in a DPA task like pair-recall neurons in IT. Also, Gutnikov et al. [3] reported that transection of the uncinate fascicle mediating direct interaction between IT and PFC leads to a leaning deficit in a PA task. Furthermore, Tomita et al. [18] demonstrated that memory retrieval in IT is controlled by a top-down signal from PFC. These findings indicate that PFC plays a crucial role in learning and performance of PA tasks.

Nevertheless, our model does not have a part corresponding to the projections from PFC to IT because we did not find sufficient computational requirements for introducing it at the cost of simplicity. For example, activity level control in networks N_1 and N_2 is important but this does not necessarily require an additional network as a controller.

In relation to this, Naya et al. [13] devised a more difficult version of the DPA task, termed the PACS (pair-association with color switch) task, and found interesting neuronal activities in IT. There exists significant computational difficulty in learning and performing this task, and we have obtained a preliminary result that the present model can be applied to the PACS task if we introduce a context signal modifying the action of some units of N_1. This may enable us to integrate the model with PFC on the basis of computational requirements.

Together with further development of the model to overcome these limitations, experimental verification of our theory also remains for future study.

Acknowledgements

We thank Y. Miyashita for helpful suggestions. This study was supported by a Grant-in-Aid for Scientific Research on Priority Areas (#08279105, #12050209, #12210038) and Special Coordination Funds for Promoting Science and Technology from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] D.J. Amit, S. Fuji, Paradigmatic working memory (attractor) cell in IT cortex, Neural Comput. 9 (1997) 1071–1092.
[2] H. Eichenbaum, Y. Otto, N.J. Cohen, Two functional components of the hippocampal memory system, Behav. Brain Sci. 17 (1994) 449–517.
[3] S.A. Gutnikov, Y. Ma, D. Gaffan, Temporo-frontal disconnection impairs visual–visual paired association learning but not configural learning in macaca monkeys, Eur. J. Neurosci. 9 (1997) 1524–1529.
[4] S. Higuchi, 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 93 (1996) 739–743.
[5] E.K. Miller, L. Li, R. Desimone, Activity of neurons in anterior inferior temporal cortex during a short-term memory task, J. Neurosci. 13 (1993) 1460–1478.
[6] E.K. Miller, R. Desimone, Parallel neuronal mechanisms for short-term memory, Science 263 (1994) 520–522.
[7] Y. Miyashita, H.S. Chang, Neuronal correlate of pictorial short-term memory in the primate temporal cortex, Nature 331 (1988) 68–70.
[8] Y. Miyashita, Neuronal correlate of visual associative long-term memory in the primate temporal cortex, Nature 335 (1988) 817–820.
[9] M. Morita, Associative memory with nonmonotone dynamics, Neural Netw. 6 (1993) 115–126.
[10] M. Morita, Memory and learning of sequential patterns by nonmonotone neural networks, Neural Netw. 9 (1996) 1477–1489.
[11] M. Morita, Computational study on the neural mechanism of sequential pattern memory, Cogn. Brain Res. 5 (1996) 137–146.
[12] E.A. Murray, D. Gaffan, M. Mishkin, Neural substrates of visual stimulus–stimulus association in rhesus monkeys, J. Neurosci. 13 (1993) 4549–4561.
[13] Y. Naya, K. Sakai, Y. Miyashita, Activity of primate inferotemporal neurons related to a sought target in pair-association task, Proc. Natl. Acad. Sci. USA 93 (1996) 2664–2669.

[14] Y. Naya, M. Yoshida, Y. Miyashita, Backward spreading of memory-retrieval signal in the primate temporal cortex, Science 291 (2001) 661–664.
[15] G. Rainer, S.C. Rao, E.K. Miller, Prospective coding for objects in primate prefrontal cortex, J. Neurosci. 19 (1999) 5493–5505.
[16] K. Sakai, Y. Miyashita, Neural organization for the long-term memory of paired association, Nature 354 (1991) 152–155.
[17] A. Suemitsu, M. Morita, A neural network model of pair-association memory in the inferotemporal cortex, in: Proceedings of the 6th International Conference on Neural Information Processing, 1999, pp. 790–794.
[18] H. Tomita, M. Ohbayashi, K. Nakahara, I. Hasegawa, Y. Miyashita, Top-down signal from prefrontal cortex in executive control of memory retrieval, Nature 401 (1999) 699–703.