A model of emergence of reward expectancy neurons using reinforcement learning and neural network

Shinya Ishii and Katsunari Shibata, Oita university
Munetaka Shidara, National Institude of Advanced Industrial Science and Technology

Abstract In an experiment of mult-trial task using a monkey in which some successful trials are required until it gets a reward, some neurons that relate to reward expectancy have been observed in the anterior-cingulate in its brain. The reward expectancy neuron is activated in each trial except for the reward trial. Therefore, it is difficult to explain the emergence of the neurons simply by reinforcement learning. In this paper, a model that consists of a recurrent neural network trained based on reinforcement learning is proposed. From the simulation of the model, it is suggested that such neurons can emerge to realize an appropriate value function in the transition period from the single-trial task to the mult-trial task.

Keyword reward expectancy, anterior cingulate, reinforcement learning, recurrent neural network

1 Introduction

It is thought that motivation and reward expectancy must be related to our action learning. Recently, in an experiment of multi-trial task using a monkey in which some successful trials are required until it gets a reward, some neurons that relate to schedule fraction, motivation and reward expectancy has been observed in the Anterior Cinglate and the Ventral Striatum in its brain by Shidara who is one of the authors[1]. These neurons belong to the loop circuit that is working when taking actions in response to important stimulus for emotion or motivation. The Anterior Cinglate is located in the frontal cortex, and has nerve fiber connections with various areas such as prefrontal cortex, supplementary motor area and limbic system, and bears an important role for motivation. On the other, the Ventral Striatum is located in the basal ganglia. It is reported that the basal ganglia relate to action learning based on reward, and some models has been already proposed to explain the reward-related motion generation based on reinforcement learning. It is easy to explain the motivation neurons whose activation becomes large as it approaches the reward. However, it is difficult to explain the reward expectancy neurons which activate in each trial except for the reward trial.

In this paper, a model that consists of a recurrent neural network trained based on the actor-critic type reinforcement learning is proposed, and the reason of emergence of the reward expectancy neurons observed in the physiological experiment is investigated by the analysis of the model during learning. Furthermore, we aim to reinforce the possibility that reinforcement learning is a main principle of learning in the brain.

2 The experiment using a monkey[1]

2.1 Task setting

This chapter explains multi-trial reward schedule task using a monkey. In the first stage, the monkey trains visual color discrimination trial as shown in Fig.1. At first, the monitor is pitch-black, and after 500ms a white bar light called visual cue is presented at the upper edge of the monitor. When the monkey touches a bar, the fixation point presented at the center of the monitor turns red. After the varying waiting period lasting between 400 and 1200ms, the target color becomes green, which instructs the monkey to release the bar. If the bar is released within 1sec after the onset of green target, the target turns blue to signal the monkey that the trial is correct, and the monkey can get juice as a reward after 250∼350ms.

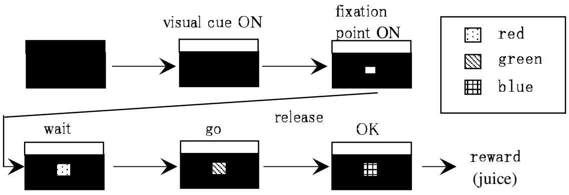


Fig.1 Visual color discrimination trial

After training of this visual color discrimination trial (single-trial task), the task transits to multi-trial reward schedule task (multi-trial task). In the multi-trial task, the reward is given to the monkey when it succeeds in successive 1∼4 trials. The necessary number of trials to get reward is determined at random. For example, the flow of the four-trial schedule is shown in Fig.2. Since the visual cue becomes bright as the monkey approaches the reward trial, it can recognize the number of trials remaining until it can get the reward. The schedule fraction, in other words, the number of trials in schedule is shown by (the number of

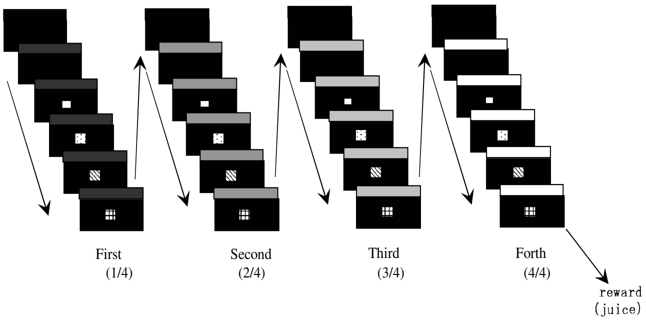


Fig.2 Multi-trial reward schedule task

trials)/(schedule). For example, 1/4 in Fig.2 indicates the first trial in the four-trial schedule. In the control experiment, the cue sequence is random not depending on the schedule so that the cue loses its meaning.

2.2 Experimental result[1]

The activation of the Anterior Cinglate neurons are shown in Fig.3. The neurons A and C in Fig.3(A)(C) generate phasic activations, while B and D in Fig.3(B)(D) keep a tonic activations. As for A and B, the activation decreased before the reward trial, and as for C and D the activation decreased after reward. It is said that A and B express the expectancy for the reward, because they did not activate in the last trial in which the reward is obtained certainly. C and D express the distance to the reward, since the activation becomes the maximum in the trial when the monkey can obtain the reward. The neurons like C and D can be explained easily as the state value(critic) by reinforcement learning. In this paper, the reason of the emergence of such neurons like A and B are investigated. Moreover, the neurons which activate only in the reward trial existing in the Ventral Striatum in the basal ganglia are focused on together with the reward expectancy neurons.

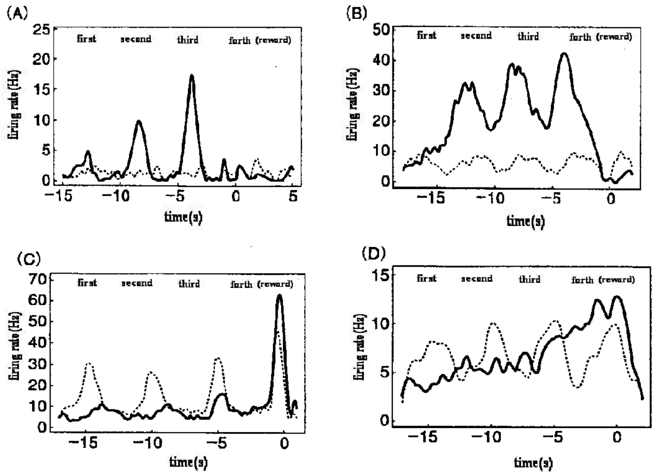


Fig.3 The response of the anterior cingulate neurons (These figures are copied from "Representation of motivational process reward expectancy in the brain", Igakuno ayumi,Vol.202 No.3,p181-p186,2002)

3 Proposed model

3.1 Usefulness of the combination of reinforcement learning and neural network

In the conventional brain research, the main purpose seems to analyze the function of each area in the brain. Also, in robotics research, each functional module such as recognition, action planning, and control was developed, and by integrating such functions, intelligent robots have been developed. In this trend, reinforcement learning has been used as the learning for the function of motion planning, and has been used in the models of the basal ganglia. However, each area in the brain is inseparably connected with each other, and it is thought that learning is done in harmony in the whole brain. One of the authors has been shown that many functions including recognition, memory and so on are acquired in a system constructed seamlessly using a neural network that is trained based on reinforcement learning. Also in each area other than basal ganglia in the brain of living things, we have thought that reinforcement learning can be a main learning principle. In the experiment of hand reaching task using a tool by a monkey, some neurons representing whether the tool is recognized as a part of the body or not are observed in the Interparietal Sulcus. The activation of the neuron representing such high order information can be explained well by the combination[6].

3.2 Proposed model

The architecture of the model proposed in this paper is shown in Fig.4. Actor-critic architecture is employed in this paper. The part called critic generates a state value from a state vector. It playes a role to evaluate the action generated by actor. Temporal difference error (TDerror) $\hat{r}_t$ is expressed by

$$\hat{r}_t = r_{t+1} + \gamma V(\mathbf{x}_{t+1}) - V(\mathbf{x}_t) \tag{1}$$

where, r: reward, $V\ (\mathbf{x}_t)$: output of the critic, $\mathbf{x}_t$:observed state, and γ: a discount factor. A neural network is used in the proposed model. The neural network is trained by the following training signals that are generated based on reinforcement learning.

$$V_{s,t} = \hat{r}_t + V(\mathbf{x}_t) = r_{t+1} + \gamma V(\mathbf{x}_{t+1}) \tag{2}$$

$$\mathbf{a}_{s,t} = \mathbf{a}(\mathbf{x}_t) + \hat{r}_t \cdot \mathbf{rnd}_t \tag{3}$$

where, $V_{s,t}$: training signal for the critic, $\mathbf{a}_{s,t}$: training signal for the actor, $\mathbf{a}(x_t)$: output of the actor, $\mathbf{rnd}_t$: trial and error factors added to $\mathbf{a}(x_t)$.

In addition, in order to deal with the past information in the neural network, a recurrent structure is introduced.

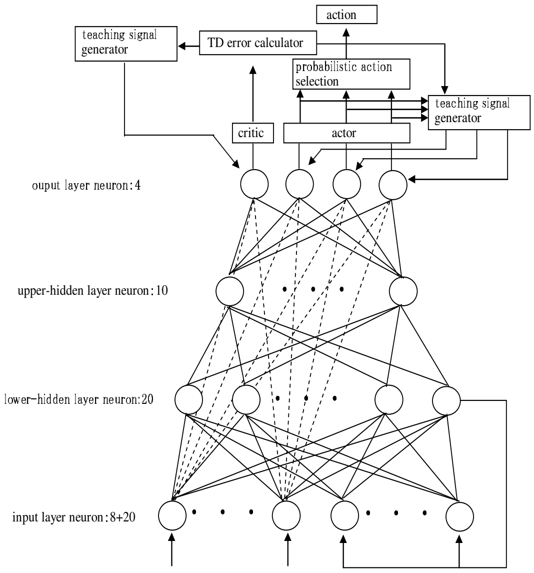

action
teaching signal generator
TD error calculator
probabilistic action selection
teaching signal generator
critic
actor
ouput layer neuron:4
upper-hidden layer neuron:10
lower-hidden layer neuron:20
input layer neuron:8+20

Fig.4 The proposed model using recurrent neural network

The number of layers is four. The number of neurons in each layer is 8 in the input layer, 20 in the lower-hidden layer, 10 in the upper-hidden layer, and 4 in the output layer. The R,G,B signal of the visual cue are inputted into the first 3 neurons in the input layer. Since the visual cue is gray scale from black to white, the values are always the same among the three neurons. The value was 1.0 in the single-trial task. In the multi-trial task, it became large as it approached to the reward such as $0.1 \rightarrow 0.4 \rightarrow 0.7 \rightarrow 1.0$. The R,G,B signals of the target color are inputted into the 4th∼6th neurons in the input layer. The signal to the input neuron 7 represents whether the monkey touched the bar or not. If the bar is touched, the signal is 1, otherwise it is 0. The signal to the input neuron 8 is 1 when the reward is given, otherwise it is 0. In addition, the direct connections from the input layer to the output layer were added. This idea is based on the knowledge that there exist different paths to the basal ganglia. One is through the frontal cortex, and the other is not through the area. It is because the latter is thought to realize an easy linear input-output relation, and the former is thought to complement the latter. Moreover, by considering that the output of each neuron is expressing pulse density, the output function of each neuron is the sigmoid function whose value ranges from 0 to 1.

The output neuron 1 is used as critic output, and the output neurons 2,3 and 4 are used as the actor output vector. One of the three actions, "keep", "touch", or "release", is assigned to each actor neuron. An action is selected statistically by comparing the values after adding a random number *rnd* to each actor output. The random number is in the range of ± 0.3. BPTT (Back Propagation Through Time) is used as a learning algorithm for the recurrent neural network, and the time to be traced back was set to 80 step. Sampling time, i.e., one step, was set to 100ms. Furthermore, when it transited from the single-trial task to the multi-trial task, the discount factor was changed as $0.96 \rightarrow 0.976$ since the necessary time steps to the reward becomes long. Each initial weight from the input layer to the lower-hidden layer or each from the lower-hidden layer to the upper-hidden layer was set to a small random value. All the weights from the upper-hidden layer to the output layer were set to 0. For this reason, at first, an action is always chosen randomly among the above three actions because the three actor outputs are the same. As for the feedback connections, self-feedback connection weights are set to 4.0, and the others to 0.0. By this setting, the propagated errors by BPTT propagate efficiently without divergence when the learning is traced back in the past, and two stable equilibrium points can be learned easily. The value 4.0 is calculated as the reciprocal of the maximum derivative of the sigmoid function.

4 Simulation rusult

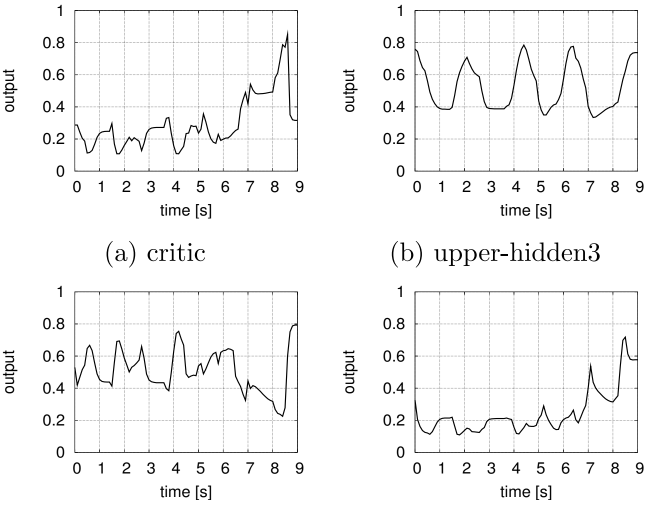
In this simulation, when the learning could be performed almost completely in the single-trial task, it moved to the multi-trial task. Here, the number of trials in the single-trial task was 16500.

4.1 The activation of each neurons

4.1.1 The result after total 20000 trials

First, the activation change of some neurons after 20000 trials, in other words, soon after switching to the multi-trial task is shown in Fig.5. The results are shown for the case when the reward is given after 4 successful trials.

The activation of the critic is shown in Fig.5(a). If the learning is performed ideally, the critic increases exponentially and smoothly toward the time when the reward is given. However, in this case, the upward trend towards the reward can be seen only in the reward trial after 6 second. Henceforth, the upper-hidden neurons 3,4 and 9, which are considered to contribute to the critic greatly by judging from the weight value to the critic, are observed. The activation of them after 20000 trials are shown in Fig.5(b)∼(d), and the change of the weight from each of the neuron to the critic are shown in Fig.6. Here, since the upper-hidden neuron 4 shown in Fig.5(c) has a negative connection to the critic, the output value is observed after turning the value upside down. In this case, the critic is expressed mainly by the upper-hidden neurons 4 and 9 in Fig.5(c),(d) in the last trial. In each trial except for the last one, since the reward cannot be got, a large negative TDerror appears. Therefore, it is thought that the activation was depressed greatly in non-rewarded trials.



(a) critic

(b) upper-hidden3

(c) upper-hidden4

(d) upper-hidden9

Fig.5 The response of some neurons after 20000 trials

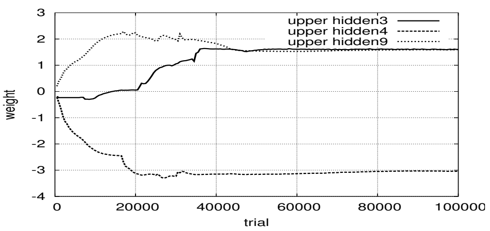


Fig.6 The change of the connection weights from the upper-hidden neuron 3,4,9 to the critic

4.1.2 The result after total 30000 trials

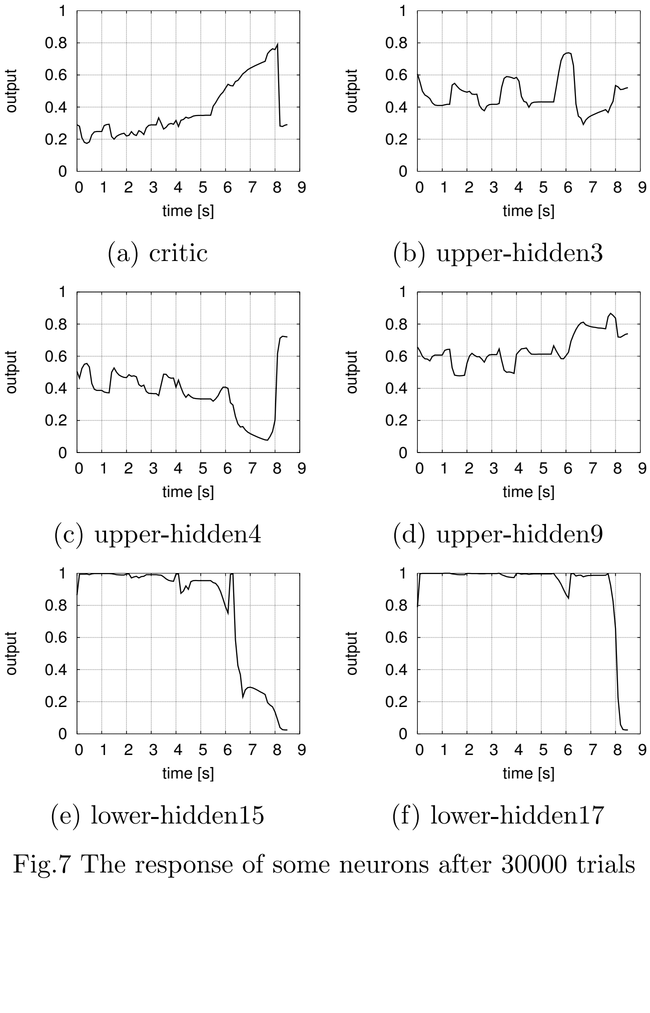
Next, the activation of each neuron after 30000 trials is shown. The activation of critic is shown in Fig.7(a). Comparing with the previous activation curve of the critic shown in Fig.5(a), the critic is increasing even before the last trial. The activation of the upper-hidden neurons are shown in Fig.7(b)∼(d). In this case, the neuron that did not activate in the last trial emerged as shown in Fig.7(b). Since the weight from the upper-hidden neuron 3 to the critic is large around 30000th trial as shown in Fig.6 as well as the upper-hidden neuron 4,9, these neurons are contributing to the critic output. Then the upper-hidden neuron 3 is considered to be equivalent to the reward expectancy neuron in the experiment using a monkey. Then, in order to consider how this reward expectancy neuron is represented, the activation of the lower-hidden neurons contributing the upper-hidden neuron3 are observed. Fig.7(e),(f) show the activations. The upper-hidden neuron3 shown in Fig.7(b) receives a positive connection from the neurons whose activation is depressed in the reward trial as shown in Fig.7(e),(f). Each of the neuron has a negative connection to the neuron which activates only in the reward trial. This means that they are contributing to both the reward expectancy neuron and the neuron which activates only in the reward trial.

4.2 Emergence reason of reward expectancy neuron

From the analysis of the above results, the reason for the emergence of the reward expectancy neuron is summarized as follows.

(1) Just after the transition to the multi-trial task, the critic is inhibited by a large negative TDerror appeared in each trial except for the reward trial. The neuron that activate only in the last trial as shown Fig.5(d) emerged.

(2) When the learning progressed to some extent, the critic output was required to increase according to the distance to the goal, and as a result, the reward expectancy neuron emerged to complement the neurons which activate only in the reward trial.



(a) critic (b) upper-hidden3

(c) upper-hidden4 (d) upper-hidden9

(e) lower-hidden15 (f) lower-hidden17

Fig.7 The response of some neurons after 30000 trials

5 Conclusion

In this paper, a model that consists of a recurrent neural network trained based on the actor-critic architecture for reinforcement learning is proposed. From the simulation of the model, a neuron that relates to ”reward expectancy” was observed in the hidden-layer. It is suggested that such neurons can emerge to realize an appropriate value function in the transition period from the single-trial task to the mult-trial task. The relation between the reason of emergence and the function of ”reward expectancy” should be considered in the future, but we think that the result supports the idea that reinforcement learning is a main principle of learning in the brain.

Acknowledgement

This research was partially supported by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research, 14350227 and 15300064.

Refarence

[1] Shidara,M. ：Representation of motivational process reward expectancy in the brain, Igakunoayumi,Vol.202 No.3, p181-p186, 2002

[2] Doya,K. ：Complementary roles of basal ganglia and cerebelum in learning and motor control, Curr. Opin. Neurobiol.Vol.10,No.6,pp.732-739, 2000

[3] Barto,A.G., Hauk,C.J., Davis,J.L. and Beiser,D.G.：Adaptive Critics and the Basal ganglia, in Models of Information processing in the basal ganglia eds.MIT Press, 1955

[4] Williams,R.J. and Zipser,D. ：Gradient-Based Learning Algorithm for Recurrent Connectioninst Networks, Tech.Report NU-CCS-90-9, Northeastern University, 1990

[5] Shibata,K. :Reinforcement Learning and Rbot Intelligence -Can Intelligence be Realized by Carrot-and-stick? - , The 16th Annual Conference of Japanese Society for Artificial Intelligence, 2002 (In Japanese)

[6] Shibata,K. and Ito,K. ：Hidden Representation after Reinforcement Learning of Hand Reaching Movement with Variable Link Length, Proc. of IJCNN, 2003, 1475-674.pdf, pp.2619-2624, 2003