Available online at www.sciencedirect.com
SCIENCE DIRECT
Neurocomputing 69 (2006) 1327–1331
NEUROCOMPUTING
www.elsevier.com/locate/neucom

A model to explain the emergence of reward expectancy neurons using reinforcement learning and neural network ☆

Shinya Ishii[a,1], Munetaka Shidara[b], Katsunari Shibata[a,*]

[a]Department of Electrical and Electronic Engineering, Oita University, Oita 870-1192, Japan
[b]Neuroscience Research Institute, National Institute of Advance Industrial Science and Technology (AIST), Japan

Available online 17 February 2006

Abstract

In an experiment of multi-trial task to obtain a reward, reward expectancy neurons, which responded only in the non-reward trials that are necessary to advance toward the reward, have been observed in the anterior cingulate cortex of monkeys. In this paper, to explain the emergence of the reward expectancy neuron in terms of reinforcement learning theory, a model that consists of a recurrent neural-network trained based on reinforcement learning is proposed. The analysis of the hidden layer neurons of the model during the learning suggests that the reward expectancy neurons emerge to realize smooth temporal increase of the state value by complementing the neuron that responds only in the reward trial.

Keywords: Reward expectancy neuron; Anterior cingulate; Reinforcement learning; Recurrent neural network

1. Introduction

Recently, in an experiment of the multi-trial task using a monkey in which several successful trials are required until it gets a reward, two types of neurons are observed in the anterior cingulate [5,4]. The one type (reward proximity type) responded more vigorously as the monkey approached the reward. The other type (reward expectancy type) increased its activity in each trial toward the reward, but dropped its activity before the reward trial. The anterior cingulate cortex is located in the medial frontal cortex. It has neural connections with various parts of frontal and limbic areas, and so can be a good candidate for integrating various information related to motivational process. The reward proximity neuron can be interpreted as the state value in reinforcement learning. On the other hand, it has been difficult to explain the emergence of the reward expectancy neuron by reinforcement learning theory because it did not respond in the reward trial. However, we thought that the reward expectancy neuron should be deeply relevant to the state value in reinforcement learning. Thus, in this paper, we hypothesized that reinforcement learning plays an important role not only in the basal ganglia, but also in the anterior cingulate cortex, and the reward expectancy neuron is one of the intermediate representations to generate state value in reinforcement learning. To investigate this hypothesis, we used a model that consists of a recurrent neural-network (RNN) trained based on the actor–critic type reinforcement learning [1], and analyzed the behavior of the hidden-layer neurons during learning. By the same approach, some neuronal responses in the intraparietal sulcus have been already explained well in terms of reinforcement learning theory [3].

2. Experimental result [5,4]

In the first stage, a monkey is trained a single visual color discrimination task [5,4]. At the beginning, a white bar called visual cue is presented at the upper edge of a black monitor. When the monkey touches a bar in the monkey chair, the fixation point presented at the center of the

☆This research was partially supported by the JSPS's Grants-in-Aid for Scientific Research (#14350227,#15300064), and also supported by AIST.
*Corresponding author.
E-mail address: shibata@cc.oita-u.ac.jp (K. Shibata).
[1]Present address: Graduate School of Comprehensive Human Science, University of Tsukuba, Tsukuba 305-8577, Japan.

0925-2312/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2005.12.101

monitor changes to the red target stimulus. After a varying waiting period, the target color becomes green, which instructs the monkey to release the bar. If the monkey releases the bar within 1 s, the target turns blue to indicate that the trial is successful, and the monkey can get juice as a reward. After the monkey learned this single-trial task, the multi-trial reward schedule task (multi-trial task) is introduced. In this task, the reward is given to the monkey when it performs 1–4 trials correctly. The necessary number of trials to get the reward is determined at random. Since the visual cue becomes brighter as the monkey approaches the reward trial, it can recognize the number of trials remaining for the reward.

The example responses of anterior cingulate neurons are shown in Fig. 1. The responses of the neuron in Fig. 1(A)(B) increased in the non-reward trials, but decreased before the reward trial. The responses of the neuron in Fig. 1(C),(D) decreased after the reward, and they can be interpreted to express the distance to the reward. The neurons like (C) and (D) can be explained reasonably as the state value by reinforcement learning. In this paper, we investigate why the reward expectancy neurons such as (A) and (B) emerged in terms of reinforcement learning theory.

3. Proposed model

The architecture of the model proposed in this paper is shown in Fig. 2. The model is consisted of one RNN whose input is an observation vector. The actor–critic [1] is employed as a reinforcement learning method. The critic output generates a state value, and the actor outputs generate action commands. From the necessity to efficiently modify all the synapse weights based on reinforcement learning even in the hidden layers as a model of the frontal cortex, error-back-propagation type supervised learning was employed, and the training signals were generated autonomously according to reinforcement learning. Thus, it is expected that necessary functions emerge purposively, autonomously and in harmony as an intermediate representation to generate appropriate state value and actions [2].

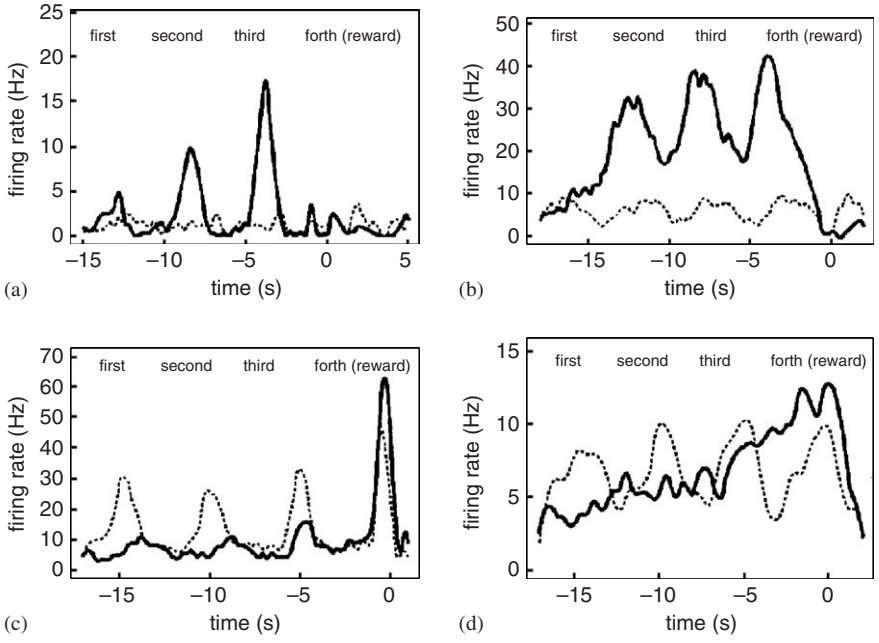

Fig. 1. The responses of the anterior cingulate neurons. These figures are copied from [4]. © 2002 by Igakushoin.

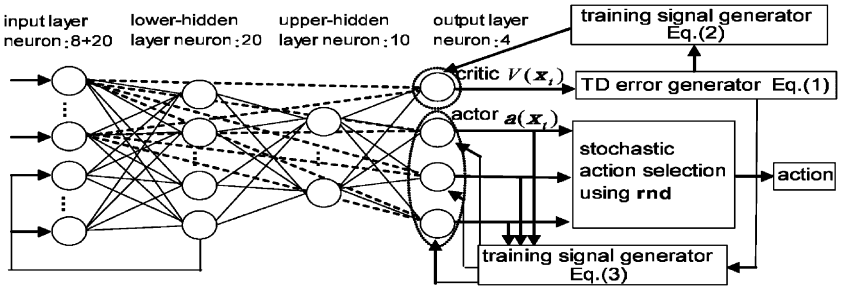

Fig. 2. The proposed model using a recurrent neural network.

TD-error $\hat{r}_t$ is expressed by

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

where r is the reward (0.9 or 0.0), $V(\mathbf{x}_t)$ is the critic output (state value), $\mathbf{x}_t$ is the observation vector, and γ is a discount factor. The neural network is trained by the training signals $V_{s,t}$ and $\mathbf{a}_{s,t}$ for the critic and actor, respectively. Those are generated based on reinforcement learning as

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

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

where $\mathbf{a}(\mathbf{x}_t)$ is the actor output vector, $\mathbf{rnd}_t$ the exploration factors added to $\mathbf{a}(\mathbf{x}_t)$.

As for the neural-network structure, the number of layers is four, and Elman-type RNN is introduced to deal with the past information. Furthermore, the direct connections from the input layer to the output layer that corresponds to the basal ganglia were added. The frontal cortex is generally considered to realize high-order functions and also short-term memory as seen in dorsolateral prefrontal cortex. Here, the information that can be approximated enough as a linear combination of the input signals is assumed to go directly from the input to the basal ganglia, while the complicated nonlinear transformation is assumed to be done by going through the frontal cortex that is modeled as the hidden layers. Since a strong nonlinear transformation is required to generate the response of reward–expectancy neurons in the anterior cingulate from the visual cue signal, at least four-layer structure is necessary after adding the output layer as the basal ganglia. The number of neurons in each layer from the input layer to the output layer is 8, 20, 10 or 4, respectively. The input signals of the neural network are assumed to be the signals after some pre-processing in the visual cortex or some other areas. The RGB signals of the visual cue are inputted into the first three input neurons, but each value is the same to the others since the visual cue is gray scale. The value was 1.0 in the single-trial task. In the multi-trial task, it became larger as it approached to the reward such as $0.1 \to 0.4 \to 0.7 \to 1.0$. The next three inputs indicate the RGB signals of the target color. The next one represents by the binary values whether the monkey touched the bar or not. The last input signal is 1 when the reward is given, and it is 0 otherwise. By considering the rate coding, the activation function of each neuron in the hidden and output layers is the sigmoid function whose value ranges from 0.0 to 1.0.

One of the output neurons is used as critic, and the other three output neurons are used as actor. One of the three actions, "keep", "touch", or "release", is assigned to each actor neuron, respectively. An action is selected stochastically by comparing the values after adding the exploring factor $\mathbf{rnd}_t$ to the actor output vector $\mathbf{a}(\mathbf{x}_t)$. Each factor of $\mathbf{rnd}_t$ is a uniform random number between ± 0.3. BPTT (Back propagation through time) [6] is used as a supervised learning algorithm for the RNN, and the truncated trace time to the past was set to 80 steps. Sampling rate, i.e., one step, was set to 100 ms. Furthermore, when the task was changed from the single-trial task to the multi-trial task, the discount factor γ was changed as $0.96 \to 0.976$ since the necessary time steps to the reward becomes large. In this simulation, when the learning was done almost completely in the single-trial task, it moved to the multi-trial task. Here, the number of episodes in the single-trial task was 16 500. An episode is defined as a sequence until the monkey gets the reward.

4. Simulation result

The response change of some neurons after 18 500 episodes, in other words, soon after switching to the multi-trial task is shown in Fig. 3. The results are shown for the case when the reward is given after 4 successful trials. The response of the critic is shown in Fig. 3(a). If the learning is performed ideally, the critic output increases exponentially and smoothly toward the time when the reward is given. However, in this case, the upward trend toward the reward can be seen only in the reward trial after 6 s. The responses of the upper-hidden neurons 3 and 9 are shown in Fig. 3(b),(c). Judging from the connection weight to the critic, the upper-hidden neuron 9 made a large contribution to the critic. In the non-reward trials, a large negative TD-error appears because the reward cannot be obtained on the contrary to the expectation. Therefore, it is thought that the critic response was depressed greatly in the non-reward trials, and that is the reason why the upper-hidden neuron 9, which is called reward-trial neuron here, responded only in the reward trial.

Next, the responses after 30,000 episodes are shown in Fig. 4. Comparing the critic output as shown in Fig. 4(a) with the previous one as shown in Fig. 3(a), it can be seen that the critic output is increasing even before the reward trial. The responses of the upper-hidden neurons are shown in Fig. 4(b),(c). In this case, the neuron that responded only in the non-reward trial emerged as shown in Fig. 4(b). The weight from the upper-hidden neuron 3 to the critic was small around the 18 500th episode, but it became large around the 30 000th episode. 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 examine how this neuron is represented, the response of the lower-hidden neuron contributing to the upper-hidden neuron 3 was observed. As shown in Fig. 4(d), the response of the lower-hidden neuron 15 is depressed in the reward trial. It had a positive connection to the upper-hidden neuron 3 and a negative connection to the upper-hidden neuron 9. From the above result, it can be thought that the reward expectancy neuron emerged to realize the smooth temporal increase of the critic output by complementing the reward trial neuron. The lower-hidden neuron that also contributes to the reward-trial neuron makes it possible for the reward expectancy neuron to generate the intermediate

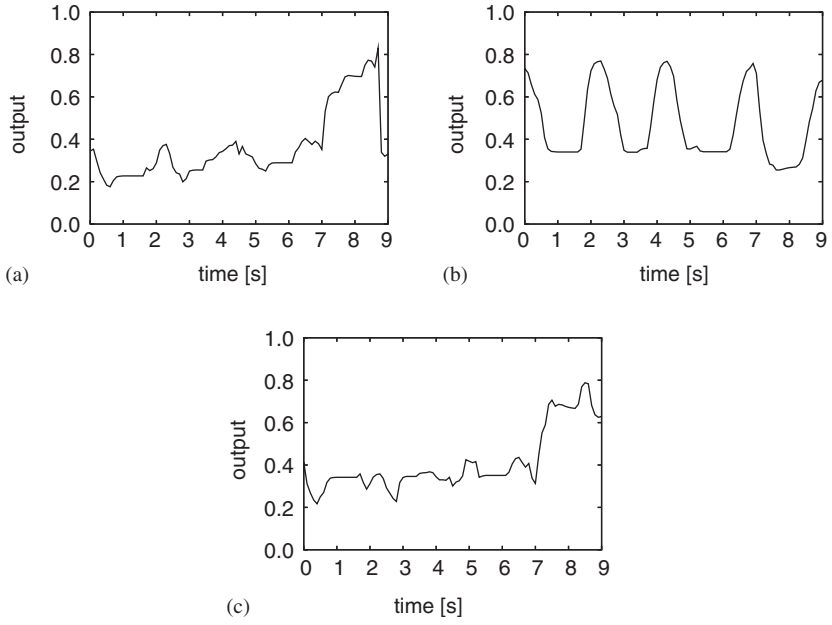

Fig. 3. The response of some neurons after 18 500 episodes. (a) critic, (b) upper-hidden3, and (c) upper-hidden9.

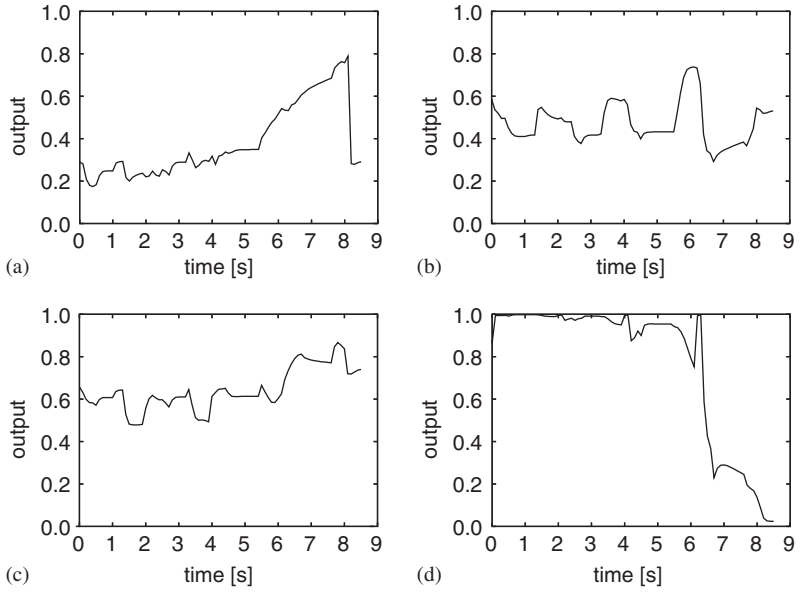

Fig. 4. The response of some neurons after 30 000 episodes. (a) critic, (b) upper-hidden3, (c) upper-hidden9, and (d) lower-hidden15.

representation in spite of a strong nonlinear transformation from the visual cue signal.

5. Conclusion

In this paper, to explain the emergence of the reward expectancy neuron in terms of reinforcement learning theory, a model that consists of a RNN-trained based on the actor–critic reinforcement learning is proposed. In the simulation of the model, a neuron that can be considered as a "reward expectancy neuron" was observed in the hidden layer. The analysis of the result suggests that it emerged to complement the neuron that responds only in the reward trial for realizing the smooth temporal increase of the critic output. The neuron responding only in the reward trial emerged to realize quite different critic output between non-reward trial and reward trial even though the input signals are similar between them, and also that makes it possible to realize the response only in the non-reward trials in the reward expectancy neuron.

References

[1] A.G. Barto, et al., Neuronlike adaptive elements that can solve difficult learning control problems, IEEE Trans. Syst. Man Cybern. 13 (1983) 835–846.
[2] K. Shibata, Reinforcement learning and robot intelligence, in: Proceedings of the 16th Annual Conference of JSAI, 2002, 2A1-05 (in Japanese).
[3] K. Shibata, K. Ito, Hidden representation after reinforcement learning of hand reaching movement with variable link length, in: Proceedings of IJCNN, 2003, pp. 2619–2624.
[4] M. Shidara, Representation of motivational process reward expectancy in the brain, Igaku No Ayumi 202 (3) (2002) 181–186 (in Japanese).
[5] M. Shidara, B.J. Richmond, Anterior cingulate: single neuronal signals related to degree of reward expectancy, Science 296 (2002) 1709–1711.
[6] R.J. Williams, D. Zipser, Gradient-based learning algorithm for recurrent connectionist networks, Technical Report, NU-CCS-90-9, Northeastern University, 1990.

Shinya Ishii is a graduate student in the Department of Electrical and Electronic Engineering, Faculty of Engineering, Oita University, Japan. He received a B.Eng. in Electrical and Electronic Engineering from Oita University in 2004. His research interests include neuronal model that consists of a recurrent neural network trained based on reinforcement learning.

Munetaka Shidara is a professor of the Institute of Basic Medical Sciences and Doctoral Program in Kansei, Behavioral and Brain Sciences, Graduate School of Comprehensive Human Sciences, University of Tsukuba in Japan. He received Ph.D. in Basic Medical Science from the University of Tokyo in 1990. His research interests include neuronal and information processing mechanisms on brain functions such as motiva tion and goal-directed behavior, and pattern recognition.

Katsunari Shibata is an associate professor of the Department of Electrical and Electronic Engineering, Faculty of Engineering at Oita University. He received a Dr. of Engineering in Electronics Engineering from the University of Tokyo in 1997. His research interests include the difference of intelligence between real lives and modern robots and also autonomous learning especially using reinforcement learning and neural network.