A conscious AI system based on recurrent neural networks applying dynamic information equilibrium

Yasuo Kinouchi[1][0000-0002-3558-0442] , Kenneth James Mackin [1] [0000-0001-9829-5254] and Pitoyo Hartono [2][0000-0002-2807-6002]

[1] Tokyo University of Information Sciences, Chiba, Japan
[2] Chukyo University, Nagoya, Japan
kinouchi@rsch.tuis.ac.jp

Abstract. A basic structure and behavior of a human-like AI system with a function equivalent to consciousness is proposed. The system is constructed completely with artificial neural networks (ANN), and an optimal-design approach is applied. The proposed system uses recurrent neural networks (RNN), which execute learning under dynamic equilibrium, instead of feed-forward ANNs in the previous system. The redesign using RNNs allows the proposed brain-like autonomous adaptive system to be more plausible as a macroscopic model of the brain. By hypothesizing that the "conscious sensation", which constitutes the basis for phenomenal consciousness, is the same as "state of system level learning", we can clearly explain consciousness from an information system perspective. This hypothesis can also comprehensively explain recurrent processing theory (RPT) and the global neuronal workspace theory (GNWT) of consciousness. The proposed structure and behavior are simple but scalable by design, and can be expanded to reproduce more complex features of the brain, leading to the realization of an AI system with a function equivalent to human-like consciousness.

Keywords: Model of consciousness, Dynamic equilibrium, Recurrent neural network.

1 Introduction

In order to realize a truly human-like AI system, it is important that the system not only models the phenomenal aspects of consciousness, but also incorporates a macroscopic model of the brain as an autonomous adaptive information system. Here, we assume that the most important function of the brain as a system is to autonomously learn and adapt itself. In the process of evolution, the brain has achieved a highly optimized structure and internal process, to improve the speed, efficiency, and effectiveness of its primary function. It is natural to assume that consciousness is an essential mechanism of the brain as an autonomous adaptive system. From this standpoint, we believe that by applying optimal or limit state design for a brain-like autonomous adaptive system, a mechanism equivalent to consciousness will inevitably become clear. For the purpose of this research, our targeted autonomous adaptive system will incorporate only the minimal functions at a very basic level, in order to clarify the basic structure and behavior of the brain as an information processing system.

In this paper, the basic structure and behavior of a human-like AI system with a function equivalent to consciousness is proposed. The system is constructed completely with artificial neural networks (ANN), and an optimal-design approach is applied, which bases the design on maximum performance and efficiency. The proposed system enables the function of consciousness to be explained from an information processing viewpoint. Kinouchi and Mackin [1] has previously proposed a conceptual structure of an autonomous adaptive system with conscious like functions. But the ANNs that constitutes the main process for autonomous adaptivity required further study and verification. In this research, the ANNs were redesigned based on the idea of dynamic equilibrium proposed by Scellie and Bengio [2].

The redesign of the ANNs allows the proposed brain-like autonomous adaptive system to be more plausible as a macroscopic model of the brain. The redesign also shows that the system can be realized by a simple structure and control method. Further, by hypothesizing that the "conscious sensation", which constitutes the basis for phenomenal consciousness, is the same as "state of system level learning", we can clearly explain consciousness from an information system perspective. This hypothesis can also comprehensively explain recurrent processing theory (RPT) [3,4]and the global neuronal workspace theory (GNWT) [5] of consciousness.

The proposed structure and behavior are simple but scalable by design, and can be expanded to reproduce more complex features of the brain, leading to the realization of an AI system with a function equivalent to human-like consciousness.

2 System Configuration and Behavior

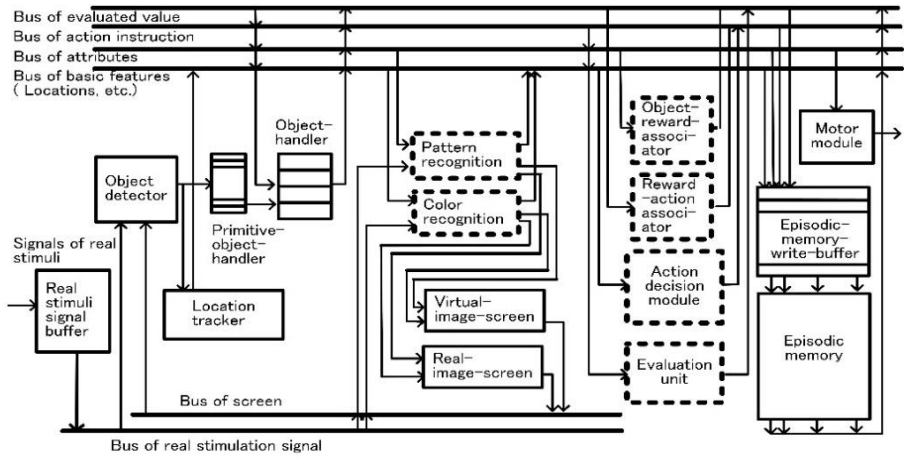

Fig. 1. Schematic configuration of the system

A. Configuration and Features

The system configuration is shown in Fig.1. The configuration follows the design by Kinouchi and Mackin [1]. The function units marked by dotted lines in Fig.1, pattern recognition unit, color recognition unit, evaluation unit, action decision module, etc.

have been constructed using recurrent neural networks (RNN), and can be trained. The other units are created with fixed functions. The general behavior of a RNN can be described as the temporal change in dynamic equilibrium of the network, or minimum energy state of the circuit, caused by the recurrent stimulation among the interconnected nodes. Scellie and Bengio uses 2 different dynamic equilibrium to control RNN. (1) Free phase: The input nodes are clamped with the input pattern signal to achieve a dynamic equilibrium. (2) Weakly clamped phase: In addition to the clamped input nodes, the output nodes are also weakly clamped to a desired output, in order to shift the dynamic equilibrium towards a desired state [2].

The RNNs are first operated in the free phase, followed by the weakly clamped phase, for the network learning. Pattern recognition is done using the free phase. Hebbian learning is used to train the weights based on the difference of activity of each node between the free phase and weakly clamped phase.

There are 2 merits of adopting this method for a brain-like autonomous adaptive system. (1) It becomes possible to train the RNN by retaining the desired state for a short period of time, regardless of the current structure or state. (2) Training the network weights using Hebbian learning allows the RNN implementation to be feasible from an engineering standpoint, as well as being a plausible model of the human brain.

B. Basic Behavior and Features

The system behavior is shown in Fig.2. The basic flow of processes follows the design by Kinouchi and Mackin [1].The basic flow of the system repeats the cycle of 1)preprocessing phase, in which object detection and pattern recognition occur, 2)decision phase, in which the system selects the most desirable object-action pair from among several detected objects, and 3)postprocessing phase, in which the system reconfigures and coordinates major information scattered in the system and executes system level learning. In the system level learning, broadcasting of system-level-shared-information, learning of related RNNs including screen depiction are concurrently processed.

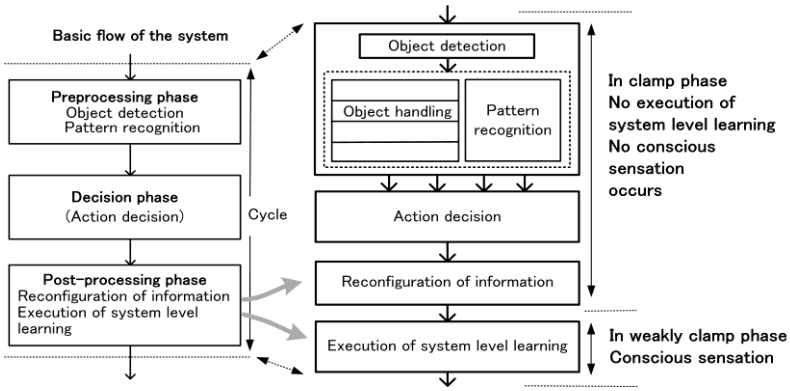

Fig. 2. Behavior of dynamic equilibrium in the system

Other than the action decision module already designed using RNN, the network structure for units with learning abilities was changed to RNN. This change only creates

a difference in system behavior during the postprocessing phase. By applying the weakly clamped phase by Scellie and Bengio [2], pattern recognition and evaluation units can be trained by simply retaining the updated states for a short period of time. The updated states are a part of system-level-shared-information, including recognized object-attribute sets, and prediction error for evaluation modules. Concurrently the retained information are sent to episodic memory to "memorize" the information through learning. The action-decision module do not train in awake-mode, but later train during sleeping mode by reading out information from the episodic memory.

Kinouchi and Mackin has already shown that learning at the system level is indispensable for an autonomous adaptive system, and that the information used for learning is equivalent to conscious sensations in phenomenal consciousness [1]. But the connection between conscious sensations and network learning was not yet clear. For the newly proposed method, the RNN in the autonomous adaptive system learns by retaining necessary information for short periods of time, suggesting that conscious sensations occur when information is retained for system level learning. The hypothesis recently proposed by Lamme that conscious sensations occur in our brain during recurrent processing to change the RNN structure itself [4], supports our proposal.

3 Consciousness in Autonomous Adaptive Systems

A. Proposed model and function of consciousness

(1) The function of conscious sensation

Conscious sensation corresponds to system level learning based on the system-level-shared-information primarily consisting of the selected target object and its evaluated value by the system. The evaluated value corresponds to our feeling of pleasant/unpleasant or comfort/discomfort, which indicates the direction of the change of configuration as a whole system. Therefore, the evaluation unit must be activated for a conscious sensation to occur. It follows that an evaluation feature is inevitable for a system to realize a function equivalent to consciousness.

(2) The function of image

We define image as "information generated inside the system that the system can operate as an object (processing target)", following the definition by Haikonen [6]. In the post-processing phase, object image is written to short-term-memory screens, the real-image-screen and the virtual-image-screen, during the RNN learning process. Object information depicted on the screen can be handled by the system as objects or processing targets in the next cycle. In the proposed system, the image-handling feature is extended so that internal information can be written back onto the virtual-image-screen as an image, using an autoencoder in the pattern recognition unit to reproduce signals, so that the system can produce a conscious sensation with or without actual external stimuli. By this feature, the system can recall information stored in episodic memory to produce a conscious sensation, and further can process this information as an object. We assume that the image on the virtual-image-screen corresponds to our mental imagery. The repeated processing of the virtual-images corresponds to how our mind "thinks".

(3) The function of self

In order to express the relationship between an object and the system from the viewpoint of the system itself, information regarding the system itself is unnecessary, and only an evaluated value expressing the relationship is required. Based on this view, we regard the state of the evaluation unit as a kind of system representative "self ". The system is at the origin of its model of the environment, and object positions are represented relative to the origin. This system-object relationship generates the sensation of the "self" seeing real objects in its environment, and constitutes the basis of the first-person perspective or the subjective experiences as if the homunculus in our brain sees the outside world as shown in Fig.3.

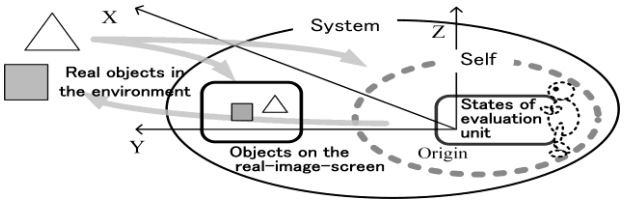

Fig. 3. A schematic diagram of a first-person perspective

B. Validity of the consciousness model

The function of consciousness must be considered through the behavior of the whole system, and cannot be defined correctly using only a limited viewpoint or section of the system. For the whole system to maximize its ability, it naturally requires all of the resources to be managed collectively each cycle. Unity, a key feature of consciousness, exists for this purpose.

Our proposed model of consciousness is consistent with both the recurrent processing theory (RPT), and the global neuronal workspace theory (GNWT) [5], two of the most potential models of consciousness. RPT has held that consciousness is associated with activity in RNN, but a new proposal suggesting that learning in RNN is a strongly connected to consciousness has been recently reported [4]. Our proposed method can be interpreted as a system-level implementation using RNN proposed by Scellie and Bengio [2], fulfilling the recent proposal in RPT. Furthermore, in our proposed method, the conscious information must be broadcast and shared within the whole system. From this viewpoint, our method is consistent with GNWT.

On the other hand, the integrated information theory (IIT) [7] is weak in its theory of consciousness, from the view that it lacks consideration of the connection between autonomous adaption and consciousness.

4 Conclusion

In this paper, we showed that it is possible to construct an AI system with a function equivalent to consciousness, by including RNNs based on dynamic equilibrium to the previous proposal by Kinouchi and Mackin [1].

We have begun the verification of the proposed system through software simulation, and are currently in the process of expanding the simulation and improving the details of the proposed model, in order to upgrade the simulation from a toy-model to a more practical conscious AI system.

We believe that conscious AI systems are advantageous in the following points.

1) A conscious system can adapt itself to dynamic environments by trial and error, based on its own goal and evaluation function. By combining with other machine learning methods including deep learning, a highly flexible and adaptive AI system that can solve problems on its own can be realized.

2) Since a conscious AI system has similar information processing characteristics with the human brain, the AI system interface will be more human-like and natural to the user.

The term artificial intelligence, coined over 60 years ago, originally targeted at artificially reproducing human intelligence on computers. But our standpoint is that the core essence of intelligence is not a unique human trait, and is not limited to the human brain. In order to clarify the true essence of intelligence, we need to redefine intelligence as a natural phenomenon. In other words, we need to consider intelligence as a form of information processing explained as a natural or physical phenomenon. Physical phenomena can be explained as different particles interacting with each other and approaching a state of dynamic equilibrium, such as a system in minimum-energy state.

RNNs can be viewed as a system of interacting elements expressed as a network structure. The network behavior is such that the network aims to achieve a stable minimum-energy state. RNNs have a large freedom in the design of interaction and energy function, and the proposed method by Scellie and Bengio [2] greatly reduced the difficulty of designing and training RNNs. We believe that by applying the idea of dynamic equilibrium to information processing in autonomous adaptive systems, we can approach the key question “what makes intelligence”.

References

1. Kinouchi,Y., Mackin,K.J.: A Basic Architecture of an Autonomous Adaptive System With Conscious-Like Function for a Humanoid Robot. Front. Robot. AI 5:30. (2018). doi: 10.3389/frobt.2018.00030
2. Scellie,B., Bengio,Y.: Equilibrium Propagation: Bridging the Gap between Energy-Based Models and Backpropagation. Front. Comput. Neurosci. 11:24. (2017). doi: 10.3389/fncom.2017.00024
3. Lamme,V.A.F.: How neuroscience will change our view on consciousness. COGNITIVE NEUROSCIENCE,1(3),204–240(2010)
4. Lamme,V.A.F.: Challenges for theories of consciousness: seeing or knowing, the missing ingredient and how to deal with panpsychism, Phil. Trans. R. Soc. B 373: 20170344.(2018)
5. Dehaene, S.,Lau,H., Kouider,S.: What is consciousness, and could machines have it?. Science 358, 486–492 27 October(2017)
6. Haikonen,P.: The Cognitive Approach to Conscious Machines. Exeter: Imprint academics (2003).
7. Tononi,G.: Integrated information theory of consciousness: an updated account. Arch. Ital. Biol. 150, 290–326(2012).