Reinterpreting The Cortical Circuit

Hiroshi Yamakawa [*1, *3], Naoya Arakawa [*1, *3], Koichi Takahashi [*2, *3, *4]
*1 Dwango Co., Ltd., *2 RIKEN Quantitative Biology Center, *3 The Whole Brain Architecture Initiative, *4 Keio University Graduate School of Media and Governance

Abstract

As various functions are accomplished with a uniform mechanism of the micro-circuit in the neocortex (i.e., the canonical cortical circuit), this system will yield clues about general intelligence. Though understanding the mechanism of the entire brain and its circuits would serve the creation of artificial general intelligence, its required specifications have not been clarified. In this paper, we present a framework for the canonical circuit, defining its functions and interface semantics while integrating models in neuroscience. In addition, candidate models with artificial neural networks are evaluated with the specifications.

1 Introduction

The whole brain architecture approach is a research approach aiming at engineering human-like artificial general intelligence (AGI) learning from the entire architecture of the brain, where major brain organs such as the neocortex, basal ganglia, hippocampus, amygdala, thalamus, and cerebellum are implemented as machine learning modules that are connected based on neuroscientific findings such as mesoscopic connectomes.

The approach aims to build AGI that acquires the ability to solve various problems through learning. The neural circuit of the neocortex has a uniform structure, but can exhibit various functions according to the nature of the input and output. Thus, the neocortex is considered to be the brain organ for versatility. In neuroscience, studies on the canonical cortical circuit/model of the neocortex have been carried out [Harris and Mrsic-Flogel, 2013]. We call the hypothetical general machine learning algorithm of the neocortical circuit the Neocortical Master Algorithm (NMA hereafter), named after the appellation for the ultimate learning machine introduced in [Domingos, 2015].

In this paper, we identify the input and output that are indispensable for NMA based on knowledge in neuroscience and machine learning, list machine learning algorithms related to NMA, and discuss its I/O and required functions.

Here, we assume cognitive architecture to operate in real time, as it is the case in the biological brain.

2 The NMA framework

With the advent of artificial neural networks (ANN) based on deep learning, there are an increasing number of systems that can realize various human cognitive abilities. These examples suggest that NMA could be realized by combining ANNs, such as the convolutional neural network (CNN) capable of information compression and pooling, LSTM with a gating mechanism, and reinforcement learning.

2.1 Reviewing the interface of NMA

What we propose here as the NMA framework are the I/O semantics for the canonical cortical circuit. In humans, the neocortex can be decomposed into functional regions, such as the primary visual cortex (V1) and supplementary motor cortex (SMA). Those regions correspond to NMA units; thus, it is important to determine the semantics of their input and output.

While the real neocortex forms a six-layer structure, here we regard it to have the five layers {L1, L2/3, L4, L5, L6} in our model. The semantics of the input and output of NMA is determined as follows. As for output signals, as a specific subtype of neurons, output signals and output semantics depend on the layer to which the subtype belongs. As for input signals, they need to be classified according to the nature of the brain organs that transmit them, as the integration of signals with multiple semantics may take place.

In the following, we integrate and merge I/O semantics from four neuroscientific sources: the Bayesian filter hypothesis [Funamizu et al., 2016], the canonical neocortex circuit by AGI.io [Kowadlo and Rawlinson, 2015] [Kowadlo and Rawlinson, 2016] the Cognitive Consilience [Solari and Stoner, 2011], and the interlaminar relationship of the neocortex [Hawkins, 2010], as shown in Fig. 1.

2.2 I/O Semantics of NMA

Here, signals are classified into state signals and control signals, where the former represent a belief in the current situation, corresponding to sensor information obtained from the external world in real time. State signals are the output from L2/3 (Output 3 in Fig.1) to L2/3 in other corteces, L4 in the higher cortex, and hippocampus. The

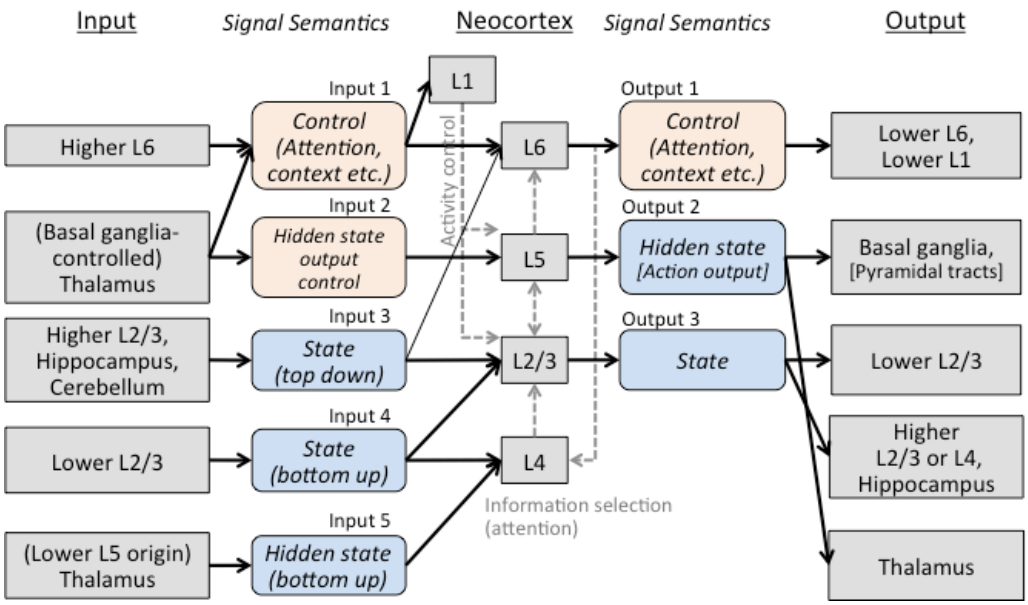

Figure 1: Framework of Neocortical Master Algorithm

"hidden state" signal is a signal with temporal information, including a history of a certain state and prediction. It is the output from L5 (Output 2) and is transmitted to the basal ganglia and thalamus. The "action output" signal is the output from another subtype of neurons in L5 to the pyramidal tract.

Of the input signals, those related to states are classified into the "state (bottom-up)" signal (Input 3) received from the lower cortex, and the "state (top-down)" signal (Input 4) received from the higher cortex. The state (top-down) signal obtained from L2/3 of the higher cortex and hippocampus is mainly the input to L2/3 and further to L6. The state (bottom-up) signal (Input 4) obtained from L2/3 of the lower cortex is the input to L4 or L2/3. The hidden state signal (Input 5) originating from L5 in the lower cortex is projected (transmitted) through the thalamus into L4. Here, note that the hidden state output from L5 carrying the information of the past is not directly sent to the lower cortex.

The "control" signal is the signal related to attention and the context outputted from L6 (Output 1) and transmitted to L6 or L1 in the lower cortex. The control signal obtained from L6 in the higher cortex or from the thalamus via the basal ganglia is the input to L1 or L6 (Input 1). The control signal obtained from the thalamus via the basal ganglia is the input to L5 as "hidden state" control output (Input 2).

In this paper, L4 is used only for information selection, such as pooling in visual information processing, the model of the external world is held in a recurrent loop between L2/3 and L5, where temporal hidden states are maintained, and the control signal from L6 controls the activity of L5 and L2/3 via L1.

There are ample neuroscientific findings for L2/3 on the recognition pathway and computational models to implement it. Findings have also been accumulating for the basal ganglia loop originating in L5, often assumed to realize reinforcement learning (see the "PBWM model" section below). As for the input and output of L6, neuroscientific findings are still scant, and we found neither a hypothesis nor a computational model about its function.

Two types of top-down signals

The top-down signal from the higher cortex to the lower cortex supports attention, prediction, and motion generation. From the viewpoint of NMA, top-down signals are categorized into two types: decoding and controlling.

The decoding signal reproduces (or predicts) representations in the lower cortex (a generation model). In Fig. 1, there is a top-down path that directly projects the state of the higher L2/3 (Output 3) to the lower L2/3 as Input 3. Input 3 also receives information from the hippocampus and cerebellum. The hippocampus may represent (index) current (or replayed) situations and cerebellum-simulated states. In this regard, they seem to send contextual information to generate representation in the cortex similar to the top-down input from Output 3.

The control signal has two origins. The control signal is transmitted top-down as Output 1 from L6 projected to L6 or L1 in the lower cortex as Input 1. The control signal is also generated in the basal ganglia, to which the hidden state (Output 2) from L5 is projected, and is sent to the neocortex as Input 1 and Input 2. As this signal is merged for various information, it cannot be classified into top-down or bottom-up.

Table 1: List of input and output and functions of existing models

		I / O semantics / functions	Ideal NMA	AGI.io	Bayesian filter hypothesis	Cognitive consillience	PredNet	HTM	BESOM	Stacked Auto Encoder	Ladder Net	CNN	Frontal cortex for PBWM
I/O		Hierarchy	✔	✔	✔	N/A	✔	TBI	✔	✔	✔	✔	N/A
	Input1	Control	✔	✔	✔	✔	_	_	_	_	_	_	✔
	Input2	Hidden state output control	✔	✔	_	✔	_	_	_	_	_	_	✔
	Input3	State (top down)	✔	✔	_	✔	✔	✔	✔	✔	✔	_	_
	Input4	State (bottom up)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
	Input5	Hidden state (bottom up)	✔	✔	✔	✔	_	✔	_	_	_	_	_
	Output1	Control	✔	✔	✔	✔	_	_	_	_	_	_	_
	Output2	Hidden state	✔	✔	✔	✔	_	✔	_	_	_	_	✔
	Output3	Control	✔	✔	_	✔	✔	✔	✔	✔	✔	✔	_
Functions	Unsupervised learning	Time series prediction	✔	N/A	N/A	N/A	✔	✔	_	_	_	_	N/A
		Disentangle / Orthogonalizat	✔				_	_	✔	_	D/J	_	
		Categorization	✔				_	D/J	✔	✔	_	_	
		Dimension reduction	✔				✔	_	✔	✔	✔	✔	
		Sparse representation	✔				✔	✔	✔	D/J	✔	_	
	Internal state	Maintain simple state	✔				✔	_	_	_	_	_	
		Finite state machine	✔				✔	✔	_	_	_	_	
		Pushdown automaton	✔				D/J	_	_	_	_	_	

D/J: Difficlult to judge, _ : Not present, ✔ Present, TBI:To be implemented

Discussion about representation of L2/3 and L5

As intelligent agents must operate based on incomplete information from the external world, the interpretation of the world becomes ambiguous. However, neural activity cannot simultaneously hold multiple states, not to mention enuerating all possibilities. Thus, in order to grasp the external world, it is necessary to interpret the current input while leaving various possibilities in view of the knowledge at hand, which is assumed to be expressed in L2/3 of the neocortex in our hypothesis.

Meanwhile, in order to work towards the outside world, it is necessary to exercise consistent interpretation from temporally spreaded viewpoints while considering the previous input from the world. Such interpretations are assumed to be represented in the L5 of the neocortex in our hypothesis.

Since these two aspects are always present, L2/3 and L5 should share information in NMA. If intelligent agents only have expressions like L2/3, they cannot form consistent intentional actions. If they only have a representation like L5, it will take actions based on biased beliefs. They can perform consistent actions while properly grasping the world by combining the states expressed in the two forms.

3 Functions of NMA

3.1 Function list

An ideal NMA will have the following functions.

Hierarchy

The learner is composed of similar or compatible computational layers through which more complex patterns are represented.

Dimension reduction

The function to approximate a high-dimensional state space in a lower-dimensional one while avoiding large information loss. It makes reinforcement learning efficient by reducing the search space to realistically occurring states.

Unsupervised Learning

Ability to change internal parameters from input sequences without teacher or reward signals. Examples include time series prediction, disentangling/orthogonalization, and categorization.

When parameters (i.e., synaptic weights) in the neocortex are determined through experience, unsupervised learning is required because teacher or reward signals are too sparse in real life.

Time series prediction

Prediction is often required for solving problems and survival. Technically, it also includes the estimation of the hidden states behind time series. As conventional techniques, the hidden-state markov model (HMM) and Kalman filter are commonly used. For prediction in the brain, there is a hypothesis called predictive coding, for which prediction error plays a key role [Rao and Ballard, 1999].

Disentangling / Orthogonalization
In a high dimensional state space, multiple factors generally exist in the form of entanglement. It becomes apparent especially when dimension reduction is performed. Disentangling or orthogonalization is the function to extract independent factors. For example, in the ATARI game task, the position (e.g., above/below/left/right) of a ball and the score can be extracted as independent variables. If independent factors are extracted, unexperienced states could be inferred from their combinations.

Categorization
The ability to classify a continuous state space into a finite discretized set. In the brain, nonparametric categorization (i.e., the number of categories are not determined) must be performed.
Categorized elements can be associated with symbols and form the basis for symbolic concepts and symbols. They also provide the discretization necessary for regarding the system as an automaton and give the foundation for action selection. They are also essential to support negative concepts and ontology [Yamakawa, 2014].

Sparse representation
Information representation for which only a small number of variables have non-zero values while the value of the majority is zero. In the neocortex, circuitry containing inhibitory cells may implement k-WTA (k-winners take all) to realize sparse representation. In engineering, regularization terms are often introduced. In a sparse representation, categorization becomes easier, because the substantive dimensions decrease as the vicinity of each state narrows, though the apparent dimensions may increase [Yamakawa, 2014].

Internal state
Internal states are essential for working memory required in various cognitive functions to hold information in an operable form. State holding is also indispensable for binding information represented in disparate locations (distributed representation in neural circuits). It is presumed to be realized with local recurrent circuits or fast synaptic plasticity. The function of simply sustaining a state like afterimage is referred to as "maintain simple state" (where the state may not be discretized) in this paper.
A system that processes time series with discrete internal states is called an automaton, categorized according to the type of recognizable formal language. For instance, a finite state machine recognizes regular grammars and a pushdown automaton (PDA) recognizes a context-free grammar with a nested structure. From the functional viewpoint, a system can be regarded as an automaton without discrete internal states as long as it handles a discrete time series. The function of automatons is assumed in living creatures. For instance, the songs of birds and whales can be represented with regular grammar, and most of the grammar of human languages can be described with context-free grammar. In order for the brain to generate and recognize time series with syntax, it must emulate the type of automaton that corresponds to the grammar. Furthermore, the ability to handle working memory can also be modeled as an automaton.
Recurrent neural networks (RNN) have been proved to be theoretically equivalent to the Turing machine, capable of emulating arbitrary automata [Siegelmann and Sontag, 1995]. In practical use, RNN is known to emulate finite state machines and PDAs [Horne *et al.*, 1998] [Gers and Schmidhuber, 2001]. RNNs with gates such as LSTM and GRU can be regarded as realizing a stack in PDA with their state-holding gate circuits. For example, LSTM recognizes and generates nested tags in XML [Graves, 2013].

Learning modulation
In recent years, global projection of neural modulators has also been suggested [Pi et al., 2013]. Such projections may control learning coefficients and reinforcement learning in the entire neocortex. Such a mechanism could be also adopted in NMA.

Parallelism
Model parallelism is one of the practical requirements for NMA to be scalable to real-world problems, eventually to the scale comparable to the brain. Many artificial neural network algorithms make use of back-propagation in the manner through which the entire network is updated synchronously. However, this synchronization bottleneck spoils the natural parallelism that biological neural circuits hold. NMA will need to be equipped with the capability to update its sub-modules concurrently and assign these computational loads to different processing units.

3.2 Reviewing existing models

In this section, we examine existing models as candidates for NMA. Table 1 compares their I/O and functions. PredNet, HTM, and BESOM are inspired by brain mechanisms and the stacked autoencoder, the ladder network, and CNN are based on engineering ideas.

Deep Predictive Coding Network (Deep PredNet)
Deep PredNet is a hierarchical neural network, with the top-down signal representing prediction and the bottom-up signal prediction error. It is developed by David Cox, et al. [Lotter et al., 2016] based on the idea of predictive coding. It performs unsupervised learning by prediction, combining CNN and LSTM to represent intralaminar coupling.

Hierarchical Temporal Memory (HTM)
HTM is a machine learning system inspired by the structure of the neocortex and characterized by prediction, sparse coding, and hierarchy (see [George and Hawkins, 2009] for neuroscientific interpretation and the white paper [Hawkins, 2010] for its implementation). While the multi-layer model may not have been implemented, data clustering has been implemented [Balasubramaniam et al., 2015].

BESOM
BESOM (BidirEctional Self-Organizing Map) is a computational model of the neocortex that combines a hierarchical Bayesian network, self-organizing maps for categorization, and an independent component analysis for orthogonalization [Ichisugi, 2007].

Stacked Autoencoder
A neural autoencoder functions as a dimension reducer, using the input and output layers with the same number of neurons and a hidden layer in-between with fewer neurons, allowing the input to be compressed in the hidden layer and the output to be decompressed. With a stacked autoencoder, end-to-end learning is often performed after pre-training from the bottom layer.

Ladder Network
A semi-supervised learning model that combines an unsupervised autoencoder with a supervised learning model that performs well in classification, even with relatively sparse training data [Rasmus et al., 2015].

Convolutional Neural Network (CNN)
A hierarchical neural network in which convolution layers for dimension reduction and pooling layers for information selection are alternately stacked, whose layers have detectors called filters. Its convolution layer is analogous to L2/3 and pooling layer to L4. It is mainly used for image processing, for which neurons in a higher layer have wider receptive fields. The input from the sensor is processed bottom-up and no loop exists in the network.

Canonical Microcircuits for Predictive Coding (CMPC)
Friston, et al., who have been formulating neural information processing in terms of the free energy principle and active inference, analyzed the structure of the neocortex from the viewpoint of predictive coding [Bastos et al., 2012]
.

PBWM model
The PBWM (Prefrontal cortex and Basal ganglia Working Memory) model proposed in [O'Reilly and Frank, 2006] has a macro-circuit consisting of the neocortex, basal ganglia, and thalamus, enabling operations that require the maintenance of states. The timing of the working memory update is determined by reinforcement learning in the basal ganglia.

4 Conclusion

In this paper, we proposed an NMA framework for general intelligence corresponding to the canonical cortical circuit. We described its I/O semantics and desirable functions. We also evaluated candidate ANN models describing the neocortical circuit and found that they neither model attentional control nor temporal hidden states, even though some of them model top-down and bottom-up communication. In this, we recognize the need for the formulation and implementation of NMA with desirable functions, toward which neuroscientists and AI researchers should enhance collaboration.

References

[Balasubramaniam *et al.*, 2015] Jahan Balasubramaniamn, C.B. Gokul Krishnaa, and Fangming Zhu. Enhancement of classifiers in htm-cla using similarity evaluation methods. *Procedia Computer Science*, 60:1516 – 1523, 2015.

[Bastos *et al.*, 2012] Andre M. Bastos, W. Martin Usrey, Rick A. Adams, George R. Mangun, Pascal Fries, and Karl J. Friston. Canonical Microcircuits for Predictive Coding. *Neuron*, 76(4):695–711, May 2012.

[Domingos, 2015] Pedro Domingos. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books. Basic Books, 2015.

[Funamizu *et al.*, 2016] Akihiro Funamizu, Bernd Kuhn, and Kenji Doya. Neural substrate of dynamic Bayesian inference in the cerebral cortex. *Nat Neurosci*, 19(12):1682– 1689, Dec 2016.

[George and Hawkins, 2009] Dileep George and Jeff Hawkins. Towards a mathematical theory of cortical micro-circuits. *PLOS Computational Biology*, 5(10):1–26, 10 2009.

[Gers and Schmidhuber, 2001] F. A. Gers and E. Schmidhuber. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, 12(6):1333–1340, Nov 2001.

[Graves, 2013] A. Graves. Generating Sequences With Recurrent Neural Networks. *arXiv e-prints*, August 2013.

[Harris and Mrsic-Flogel, 2013] Kenneth D. Harris and Thomas D. Mrsic-Flogel. Cortical connectivity and sensory coding. *Nature*, 503(7474):51–58, Nov 2013.

[Hawkins, 2010] Donna Dubinsky Jeff Hawkins. Hierarchical temporal memory including HTM cortical learning algorithms. Technical report, 2010.

[Horne *et al.*, 1998] Bill G. Horne, C. Lee Giles, Pete C. Collingwood, School Of Computing, Man Sci, Peter Tino, and Peter Tino. Finite state machines and recurrent neural networks – automata and dynamical systems approaches. In *Neural Networks and Pattern Recognition*, pages 171– 220. Academic Press, 1998.

[Ichisugi, 2007] Yuuji Ichisugi. A cerebral cortex model that self-organizes conditional probability tables and executes belief propagation. In *2007 International Joint Conference on Neural Networks*, pages 178–183, Aug 2007.

[Kowadlo and Rawlinson, 2015] Gideon Kowadlo and David Rawlinson. How to build a General Intelligence: Circuits and Pathways. Project AGI (Blog), http://agi.io/. 2015.

[Kowadlo and Rawlinson, 2016] Gideon Kowadlo and David Rawlinson. How to build a General Intelligence: An interpretation of the biology. Project AGI (Blog), http://agi.io/, 2016.

[Lotter *et al.*, 2016] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *CoRR*, abs/1605.08104, 2016.

[Markman and Gentner, 2000] Arthur B. Markman and Dedre Gentner. Structure mapping in the comparison process. *The American Journal of Psychology*, 113(4):501–538, 2000.

[O'Reilly and Frank, 2006] Randall C. O'Reilly and Michael J. Frank. Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput.*, 18(2):283–328, February 2006.

[Pi *et al.*, 2013] Hyun-Jae Pi, Balazs Hangya, Duda Kvitsiani, Joshua I Sanders, Z Josh Huang, and Adam Kepecs. Cortical interneurons that specialize in disinhibitory con- trol. *Nature*, 503(7477):521–524, Nov 2013.

[Rao and Ballard, 1999] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci*, 2(1):79–87, jan 1999.

[Rasmus *et al.*, 2015] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder network. *CoRR*, abs/1507.02672, 2015.

[Siegelmann and Sontag, 1995] H.T. Siegelmann and E.D. Sontag. On the computational power of neural nets. *Journal of Computer and System Sciences*, 50(1):132 – 150, 1995.

[Solari and Stoner, 2011] Soren Van Hout Solari and Rich Stoner. Cognitive Consilience: Primate Non-Primary Neuroanatomical Circuits Underlying Cognition. *Frontiers in Neuroanatomy*, 5:65, Dec 2011.

[Yamakawa, 2014] Hiroshi Yamakawa. Column Structures on the Neocortex as a Basis of Symbolic Representation. JSAI 2014, 2C4-OS-22a-4, 2014. (In Japanese)