Brain-Like Emergent Spatial Processing

Juyang Weng *Fellow, IEEE* and Matthew Luciw *Member, IEEE*

***Abstract*—This is a theoretical, modeling, and algorithmic paper about the *spatial* aspect of brain-like information processing, modeled by the Developmental Network (DN) model. The new brain architecture allows the external environment (including teachers) to interact with the sensory ends S and the motor ends M of the skull-closed brain B through development. It does not allow the human programmer to hand-pick extra-body concepts or to handcraft the concept boundaries inside the brain B. Mathematically, the brain spatial processing performs real-time mapping from $S(t)\times B(t)\times M(t)$ to $S(t+1)\times B(t+1)\times M(t+1)$, through network updates, where the contents of S, B, M all emerge from experience. Using its limited resource, the brain does increasingly better through experience. A new principle is that the effector ends in M serve as hubs for concept learning and abstraction. The effector ends B serve also as input and the sensory ends S serve also as output. As DN embodiments, the Where-What Networks (WWNs) present three major function novelties — new concept abstraction, concept as emergent goals, and goal-directed perception. The WWN series appears to be the first general purpose *emergent systems* for detecting and recognizing multiple objects in complex backgrounds. Among others, the most significant new mechanism is general-purpose top-down attention.**

***Index Terms*—Mental architecture, cortical representation, attention, perception, cognition, behavior, computer vision, text understanding, reasoning, regression, complexity.**

I. Introduction

WHILE a child incrementally learns new skills in an open-ended fashion, one after another, how does the child's brain learn to represent, and think about, its external world *without* a need for an internal central controller? It is clear that the learning must be "skull-closed". A "skull-closed" brain or network can only interact with the external environment through its sensory port and effector port. The main effector ports for the brain are muscles and glands. Conventionally, effector ports are often called motor ports.

A. Symbolic Representations: Skull Open

Using symbolic representations corresponds to assigning the human programmer to the role of an all-aware central controller. Cognitive Science [4], [81] and Artificial Intelligence (AI) [69], [54], [71] employ symbolic representations to model cognitive systems, assuming a restricted domain. Symbolic models assume that there is a one-to-one correspondence between each Handpicked symbol and meanings in the minds of human domain experts. Each symbol has only a unique meaning (or a unique set of meanings) and each meaning has a unique symbol (or a unique set of symbols). Because of this symbol use, cognitive system modelers use "skull-open" approaches — the central controller is in the outside human designer, who defines each internal module using a symbolic meaning, designs or train each module separately, and then manually links related modules.

A "skull-open" approach might use handpicked symbols for the extra-body concepts to assign roles for a brain region. For example, a "V1" module is designed to detect oriented edges, and an "hippocampus" module is designed to detect "Jennifer Aniston", "Mother Teresa", and the Pythagorean theorem $a^2 + b^2 = c^2$ (see Koch 2011 [42]). However, as we will see below from the Developmental Networks (DN), no internal neuron alone represents the pure meaning of any extra-body concept: (1) the firing of an internal neuron does not surely report the presence of an extra-body concept; (2) many other internal neurons fire in the presence of the same extra-body concept; (3) furthermore, the firing of an internal neuron depends on not only its bottom-up match (sensory feature), but also its top-down match (e.g., top-down attention), its lateral match, and its competition with many other neurons.

In principle, such handcrafting symbols is inconsistent with autonomous development of brain-mind functions, as argued by Weng et al. [100], and is intractable for muddy tasks discussed in Weng 2009 [89]. It is a root reason for system's high brittleness — the resulting systems are brittle in dealing with real environments over their lifetimes, since the task-aware human controller has already left the system but the system itself cannot guarantee that the human controller's domain restrictions are all met during all its operations. Weng & Chen 2000 [95] argued that this is the absence of an *applicability checker* with symbolic approaches. Such systems are not able to learn (develop) autonomously from the real physical world, with or without human supervision. The contents of any set of task-specific, handcrafted symbolic concepts are static, and so are the boundaries between the concept modules.

The high brittleness of symbolic systems seems mainly due to the static symbolic design, not primarily due to, as often argued, a lack of full domain ontology which is interpreted in AI as formal representation of knowledge as a set of concepts within a domain, and the relationships between those concepts. A human child or even adult does not have such a full domain ontology, but he is not that brittle.

B. Emergent Representations: Skull Closed

In contrast, using a "skull-closed" autonomous development approach, the developmental program (DP) that implicitly

Juyang Weng is with the Department of Computer Science Engineering, the MSU Cognitive Science Program and the MSU Neuroscience Program, Michigan State University, East Lansing, MI, 48824 USA (email: weng@cse.msu.edu). Matthew Luciw is with the Dalle Molle Institute for Artificial Intelligence (IDSIA), Manno-Lugano, Switzerland (e-mail: luciwmat@gmail.com). JW: conceptualized and drafted the paper. ML: Did the experiments and produced the data in Fig. 4 and the lower part of Fig. 5 when he was at Michigan State University. The authors would like to thank Zhengping Ji, Mojtaba Solgi and Kajal Miyan for their contributions to the experiments in the related publications cited in this paper.

regulates internal self-organization is task-nonspecific and autonomous. Such a developmental network does not need an applicability checker. A developmental agent is not perfect in a new environment or for a new task but it autonomously learns to become better for what it does until all its limited resource has been optimized for what it does in its recent life.

The brain is, of course, skull-closed during learning — its internal representations are not directly manipulatable by external human teachers. It has been proposed that each developing brain learns through grounded, autonomous interactions with the environment [61], [10], [93], [21], [63], [100]. But what regulates the brain's development and what characterizes the development? How does the brain generate and self-organize internal representations and operate fully autonomously with its skull fully closed? How does the brain learn concepts one after another incrementally, including concepts that the ancestors are not aware of? The theory about the spatial aspects of brain information processing presented here provides some attempts, although the theory does not mean to completely answer these questions.

C. Functional Novelty of the Paper

This work presents a general-purpose architecture and model for brain-like *spatial* processing, with its major advances summarized in Table I, where the major new contributions are listed in the last three columns. The temporal aspect of this brain-like processing model will appear elsewhere. Developmental Networks (DN), also called Epigenetic Developer (ED) networks, were proposed by Weng 2010 [90] as a simplified brain-mind model that incorporates 5 concepts chunks. The Where-What Network (WWN) [39] will be presented as an example of DN, and in WWN a motor zone (type motor area TM) is used to teach a *type* concept and another motor zone (location motor area LM) is used to teach a *location* concept. However, any other practically learnable concepts can be taught to a DN, at least in principle.

The new contributions of this work are as follows:

New-concept abstraction: *Concept abstraction* in spatial processing corresponds to the extraction of (abstract) concept independent of particular instances. However, since each instance is associated with particular values of multiple concepts (e.g., a particular type and a particular location), abstraction for concept C (e.g., type concept) should be invariant to all other related concepts (e.g., location), at least for all cases where these concepts are applicable. For example, assume a WWN has learned two concept categories, location and type. Location concept abstraction in output means that the outputs from the location motor area of WWN have a certain location specificity with invariance to all types. Type concept abstraction in output means that the outputs from the type motor area of WWN have a certain type specificity with invariance to all locations.

Now, for new concepts to emerge and reach this level of abstraction requires these three conjunctive conditions: (1) Each possible concept is not handcrafted during the programming time, and therefore a symbolic representation is ruled out. (2) The teachers tolerate certain variability in the motor port (e.g., an action — raising the index finger — can be slightly different each time as long as other humans can agree which finger is raised). (3) A mechanism for invariance of the new concept to all other related concepts (e.g., the type concept is invariant to other concepts, such as location, viewing angle, object scale [78], etc.), subject to the richness of experience. Earlier networks, such as Cresceptron [94], do not have general-purpose mechanisms to deal with both specificity and invariance because it uses a built-in invariance (e.g., the built-in shift invariance in Cresceptron means that it cannot learn to report location distortions such as facial expressions).

Concepts as emerging goals: The term "goal" in this paper is a task-specific term as a special case of a general brain mechanism — spatiotemporal context. Recall that the DP of the brain is not task-specific. However, "goal" is a task-specific concept of the external environment. Thus, "goal" must not be part of the DP. A goal is a spatiotemporal context that emerges in the motor area and is used as a top-down signal (e.g., to control the next focus of attention in WWN). Another term might fit a particular task better, such as intent, objective, action, target, type-to-search-for, location-to-attend, cognitive-state, or action-state. Interestingly, human language has these rich terms but they seem to correspond to the same top-down brain mechanism.

"Concepts as emerging goals" requires at least six conjunctive conditions: (1) New-concept abstraction. (2) Top-down: effectively influence the operation of all the related elements in earlier processing streams. (3) Rich concept possibility, since a goal should represent a goal of any learnable practical task. (4) Emergent. The goal is emergent from interactive experience, instead of constantly imposed (as with the IBM Deep Blue [35] for computer chess or IBM Watson for the Jeopardy game). (5) Dynamic: change in real time depending on real-time experience. (6) For all long-term, short-term and immediate goals, a goal emerges in the motor area that causes the brain to recursively recall the next immediate goal in the motor area. For an short-term goal (e.g., reach an apple), the next goal can be more immediate (e.g., lift the arm) or less immediate (e.g., to eat the apple). For a long-term goal (e.g., get PhD), the next action can also be immediate (e.g., concentrating on reading this paper) or less immediate (e.g., to get a more interesting job). Existing networks (e.g., [94], [36], [34]) do not have a clear way to dynamically emerge goals. Cresceptron 1997 [94], although having a separate (not concurrent) top-down computation pass, cannot generate a location goal because of its built-in location invariance.

Goal-directed perception requires three conjunctive conditions: (1) Goal emergence. (2) Preference: The goal biases the corresponding perception of the sensory input so that the sensory elements that are correlated with the current goal are more likely to contribute to the next action. (3) Relativeness: the sensory elements not correlated with the goal are relatively suppressed. Existing self-organizing networks (e.g., [94], [65]) do not have a clear way to deal with goal-directed perception for sensory inputs that have highly structured but irrelevant parts (e.g., natural but complex backgrounds).

TABLE I
COMPARISON AMONG SPATIAL METHODS

Issues addressed	Representation	Incremental learning	Goal-directed search	New-concept abstraction	Concepts as emerging goals	Goal-directed perception
Symbolic (Markov field, Bayesian net, etc.)	Handcraft	No	Yes	No	No	No
Prior emergent (neural net) models	Emergent	Yes	No	No	No	No
DN, WWN (also brain)	Emergent	Yes	Yes	Yes	Yes	Yes

D. Mechanistic Novelty of the Paper

Now, the mechanistic novelties of this paper are summarized.

First, it was found that, not only the sensory ends, but also the effector ends of a brain must also be memory query ports. This new mechanism is inspired by neural anatomy of the brain, well known in neuroscience, but has been challenging for researchers in computational neuroscience, cognitive science, neural networks, computer science, and electrical engineering to understand. This challenge has been great, since there is a lack of understanding about spatial representations in the brain.

Second, a new theory is reported here about a brain's internal spatial representation — the hextuple representation. In particular, a neuron in the brain has not only a receptive field and a projective field, but also 6 fields since we must consider (1) motor area as also input, (2) the sensory area as also output, (3) other neurons in the same areas as both input and output. The new, cortex-inspired hextuple representation introduced here indicates that the "much in-between" — internal representations — integrates bottom-up (sensory), lateral (other features), and top-down (motor) inputs. Furthermore, the integration is not static.

Third, the internal active representations at any time are:

- emergent: generative from experience;
- selective: linked with related components only;
- dynamic: quickly switch which neurons are active;
- distributed: no neuron alone represents any pure meaning of the extra-body concepts; and
- interactive: almost no neuron is irrelevant to top-down processing based on action (e.g., verbal, manipulatory, internal attention, internal glands)

Forth, the internal representations can be mathematically rigorously modeled and understood. We establish (1) the perfect motor output theorem, (2) the square-like tiling theorem, and the top-down effect theorem. We also show that symbolic representation is intractable — the number of required symbols is exponential in the number of concepts.

The remainder of this paper is organized as follows. Informed by the literature in biology, neuroscience, psychology, Section II provides new, integrated insights into the brain-mind. Section III presents the General-purpose architecture and the representation of cortex like processing. Section IV discusses the experiments with a variety of different networks using the theory and method presented here. A comparative discussion about existing models is presented in V. Section VI concludes with discussion.

II. PERSPECTIVES FOR BRAIN-MIND

Mind is what the brain does. The hyphenated word brain-mind stresses tight integration of the brain and the mind. In this section, we provide some perspectives from biology, neuroscience, and psychology, which motivated the work presented here.

A. The Brain is First a Developer: Cell-Centered

The brain is gradually developed with the body from conception (a single cell called zygote), to fetus, to birth, to infancy, to adulthood, through active sensorimotor experiences. This developmental process depends heavily on interactions among neighboring cells as well as the locally and remotely connected cells, while the entire process is regulated by the genome in the nucleus of every cell [62], [29].

Cell-autonomous interactions determine the representations and functions of different areas of the brain [8], [14], [83], [5]. The brain's internal representation is not totally innate. It is a compounding result of the genome and the experience. Thus, instead of modeling the extremely complex brain (end result of development) directly, it is more systematic and more tractable to model the functional equivalence of its developmental program (i.e., genome) and the process of autonomous brain development, as researchers have argued [21], [94], [63], [2], [100], [96].

Then, how can cell-autonomous interactions take place? The genetic equivalence principle [62] indicates that the genetic information in the nucleus of every cell is sufficient to regulate the development of a single cell into an adult having 100 trillion of cells and the brain with 100 billion cells. Weng et al. [99] proposed that this principle indicates that development is cell-centered — every cell is autonomous through the development, while interacting with its environment formed by other cells and the external environment.

This cell-centered development includes the body development (development from a single cell to an adult body) and brain-mind development (connections and modifications of synapses in the brain). By AMD, we mainly concentrate on the latter — brain-mind development — although the former is closely related.

Weng et al. [99] further argue that this cell-centered property of development further implies that learning is in-place — every cell is responsible for its learning all by itself and there is no extra-cellular learning mechanism that is dedicated to the cell. As explained in [99], this "in-place" concept is more precise than the conventional concept "local" in Euclidean distance because a neuronal connection is not local (can be as long as over a meter), and "local" does not mean that the learning model does not require an extra-cellar mechanism

(e.g., taking partial derivative for the input of each neuron as required by error back-propagation networks). For example, there is no extra-cellar mechanism to compute the correlation matrix of the input vector of each cell.

B. Implication of Genomic Equivalence

This cell-centered property of biological development discussed above is supported further by a deeper property of the genome:

The genomic equivalence principle [62] (dramatically demonstrated by mammal cloning) has shown that the genome in the nucleus of any somatic cell (i.e., any cell that forms the body) carries the complete genetic information for normal development (in typical ecological environments) from a single cell to an adult body having 100 trillion cells. This suggests that the basic, autonomous units of the brain development (and learning) are individual cells. Based on its genetic and cellular properties, each cell interacts with other cells during its development, including mitosis, differentiation, migration, growth of dendrites and axons, connection, synaptic adaptation and response generation. Many brain cells autonomously interact to form tissues, cortex, and areas [62], [83]. Thus, for the DP of DN, we should focus on cell mechanisms. Other multi-cell properties and capabilities are emergent.

The genomic equivalence principle implies that indeed there is no central controller for brain computation or brain learning.

For artificial developmental agents, by "*without* an internal central controller", we mean that no human programmer after knowing and understanding each task to be taught, is permitted to get into the "skull" to handcraft task-specific representations (e.g., symbolic representations or task-specific connectionist representations).

C. Biological vs. Behavioral Evidence

In addition to the above architectural implication, biology and neuroscience of development can provide detailed information about development, complementary to behavior studies.

For example, studies in developmental psychology have provided convincing evidence that the development of visual capabilities require extensive experience [10], [29], [12], [18], [28]. However, many psychological studies rely on observing the stimuli dependent behaviors from humans and higher animals, instead of directly studying the biological mechanisms inside the brain. Therefore, results from such studies are not sufficient for understanding the brain-mind because of their lack of modeling the causal processes inside the brain.

For example, based on behavior studies Susan Carey 2011 [11] interpreted that "a representational capacity is innate." Such a statement lacks a deeper biological account, giving an illusion that the zygote totally determines the new born brain.

Biologically, all phenotypes, except the zygote that defines the new life, seem not totally innate, since they are all emergent from the first cell zygote and dependent on experience (e.g., bad experience can stop the first mitosis from the zygote). Carey's interpretation is inconsistent with mounting evidence in developmental neuroscience: Prenatally in the womb, a vast amount of sensory (spontaneous and from the womb) and motor experience (e.g., kicking in the womb) is necessary for the brain to wire and generate signal-statistics dependent representation at birth (see, e.g., [47], [26], [57], [30]). The same is also true after the birth as soon as the baby opens his eyes.

D. Conventional Networks and Lack Thereof

One may doubt that simulating biological and neuroscience phenomena is sufficient to give rise to complex mind. This is indeed a major question that the conventional neural networks need to answer.

Although many basic models for artificial neural networks are inspired by neuron-like computations and use emergent representations, a large gap exists: Existing purely numeric connectionist approaches (neural networks) are deficient in their abilities to abstract well as correctly criticized 20 years ago by Minsky 1991 [54], also stated by Michael I. Jordan at the David Rumelhart Memorial Plenary Talk IJCNN 2011.

For example, conventional artificial neural networks are not well suited for *goal-directed search* (which symbolic methods do using handcrafted symbols) and *goal-directed perception* (which symbolic methods do using handcrafted object models, 3D or appearance based, in almost all published pattern recognition and computer vision methods). Minsky [54] argued that connectionist approaches are bottom-up (e.g., from pixels, in fact, grounded) and symbolic approaches are top-down (in fact, from human handcrafted abstract concepts, not machine abstracted). Between the concrete (e.g., an edge or an edge grouping) and the abstract (e.g., goal and decision), "much in-between" is missing. This paper intends to show that this "much in-between" is something like the autonomously developed internal (i.e., inside the brain skull) representations inside the WWN.

E. Symbols vs. Grounding

Before we deal with the network abstraction issue, let us consider why we argue that symbols are abstracted by humans.

From the cortex-like processing architecture and representation below, we will see that true meanings of the physical world lie in the spatiotemporal association of sensory and effector information experienced by a grounded developmental agent (brain or network), not necessarily in any computerized symbol.

Symbols are invented by humans to communicate among them, hand-written, said, typed, and manually signed. None of them has a form like a computerized symbol (e.g., the ASCII code of a word).

In fact, any form of a symbol is meaningless unless a human senses it in a grounded way (e.g., sees the visual image of a hand-written word or hears the sound of the word) whose sensation well matches his own grounded and memorized sensorimotor experience in the past.

A computer system that takes symbolic inputs (e.g., the ASCII code) is not grounded in the physical world without a human between it and the physical world. For example, it requires the human to detect, attend, and recognize a

foreground object from a cluttered and complex background scene and provide the ASCII code of the object. Likewise, a symbolic output from a computer system is meaningless without a human reading it. Thus, a system that takes only symbolic inputs and outputs is not able to directly interact with the physical world without a human in between.

Therefore, an Autonomous Mental Development (AMD) system must directly interact with the physical world to develop, if the development is truly autonomous — without a human in between. A simulated physical world is fine for simulation but one must be aware that a simulated physical world is not exactly the same as the real physical world.

F. Concepts

Then, how can a neural network abstract in a grounded way — in the physical world? In order words, how does physics give rise to abstract concepts? We need to first discuss what a concept is.

By the definition of the Merriam-Webster dictionary, "concept is an abstract or generic idea generalized from particular instances". The term "concept" in this paper refers to any concept that can be practically learned by a human. "Abstract", by definition, means "disassociated with any specific sensed instance". For example, "car" is a concept, but the "car" concept is disassociated with any particular kind of car, sedan, SUV, van, etc.

Since a concept can refer to different "levels" of meanings, almost any word and phrases in a human language can be a concept.

To give a sense of how varied a concept can be, here are 22 examples of concepts: object type, horizontal direction, vertical direction, apparent scale on the retina, distance, physical size, viewing angle, material, weight, surface texture, surface color, surface reflectance, temperature, deformability, lighting direction, lighting color, lighting uniformity, usage, purpose, ownership, price, horizontal relationship between two entities.

Each concept has concept values which themselves can be concepts. For example, task is a concept; while reading, eating, and sleeping are its three values (or subtasks, but subtask is a concept). Features (e.g., need humans) and properties (e.g., noble) involved in all tasks are also concepts. Among task, feature and property, it is hard to say which is "high" and which is "low". Therefore, concepts are not necessarily always hierarchical.

To model autonomous development, a human programmer should not handcraft concepts and their relationships. Concepts and their relationships emerge in the brain through interactions between the brain and its external environments (body and extra-body environment).

G. Representation of Concepts

Any human communicable concept is explained either through verbal actions (i.e., say it) or limb actions (e.g., write it down, or sign it in the American Sign Language).

Some cognitive scientists (e.g., Reber et al. 1980 [66] and Sun et al. 2005 [80]) believe that humans have two types of concepts (or knowledge), (1) explicit, declarative, or verbal (those for which we have a clear language term, such as type "cat") and (2) implicit, procedural, or nonverbal. (those we do not have a clear language term, such as skills of pointing to an object, dancing or singing). We propose that they correspond to two types of effector behaviors: (1) Explicit concepts correspond to actions for which human have developed clear linguistic terms. (2) Implicit concepts correspond to actions for which human have not developed clear linguistic terms (e.g., subtle moves during dancing). Therefore, the studies of those cognitive scientists are consistent with our proposal:

> Any concept can be represented at, and communicated through, the exposed effector end.

This proposal does not mean that internal representations do not represent concepts. They do. However, they are secondary for concepts, since they are emergent from the sensory ends and the effector ends (including prenatal development). In our DN theory below, we will see that internal representations are not necessarily more "abstract" than the representation at the effector ends.

H. Conceptual Perspectives

With the above perspectives, we are ready to discuss a few details.

First, there are no input and output symbols for the brain. We note that the brain does not input and output any computer symbol used in our symbolic cognitive science and AI systems, since the brain cannot guarantee the one-to-one correspondence. For example, all retina images of an object are different and all utterances of a word have different waveforms. Similarly, a brain produces muscle images, where each action is represented by at least slightly different muscle images each time. Such variation is also assumed for WWNs.

Second, effectors are subject to calibration. Although the brain is skull-closed, its sensory ends (e.g., retina, cochlea, and their subparts) and effector ends (e.g., muscles, glands and their subparts) are open to the brain's external environment (outside the brain, including the body). The teacher calibrates the actions from WWN through interactions.

Third, a numeric effector sequence can represent all practically possible brain outputs (muscles and glands). As we discussed, cognitive scientists argued that there are two types of memory, skills, and learning. The first type is explicit. The second type is implicit. However, the distinction between these two types is superficial if we consider muscles. For example, the muscle *implicit procedure* in the vocal track can pronounce any *explicit declarative* words. Therefore, there is in fact no fundamental difference between the above two types as the brain is concerned: They are all muscle contractions. Therefore, the WWN model here is general purpose, since it can drive artificial muscles and glands.

Fourth, the existing dorsal and ventral account misses their functional causality — motor. Since Mishkin et al. 1983 [55] discovered, through their brain lesion studies, that the dorsal stream and ventral stream are correlated to, respectively, space ("where") and object ("what"), Goodale & Milner 1992 [31] further refined to "how" and "what". Some experimental studies [25], [16], [23] reported and modeled the connections

of these two pathways. However, the dorsal and ventral streams experiments and models missed a major representational causality — motor signals from and to the frontal cortex [13, e.g., Fig. 7], as modeled, as far as we know first, by the WWN-1 work of Ji & Weng 2008 [39]. The WWN scheme goes beyond the traditional ventral and dorsal streams by including the motor areas — which are one of the two causalities (sensor and effector) of internal representations as explained below.

Fifth, behaviors are primary for the brain. Under the pressure of evolutionary competition, at any stage of development, the brain, with a limited resource, must produce in a timely manner context-aware behaviors that are aligned with its developmental stage. Thus, generating competition-required behaviors is the primary purpose of brain's internal representations, not for easier human programmer's understanding or intuitive computer visualization. For example, the brain does not seem to build an internal symbolic representation although it allows intuitive computer visualization.

Sixth, a basic hypothesis, waiting for further biological verifications, of this theory is that the basic biological DP mechanisms are similar across a wide variety of brains — from mammals to humans — and across different parts of the brain. Different brains are not the same and different brain areas are not same; but this does not mean that their developmental mechanisms are very different. Such similarities across different brain areas are well accepted in neuroscience (e.g., different areas in the cerebral cortex all have the similar 6-layer laminar structure [40]). The differences in brain sizes, bodies, and environments seem to play a major role in functional differences from mammals to humans. Computationally, this hypothesis is supported by two aspects, representation and completeness. In terms of representation, Weng 2012 [92] argued that any partition of brain areas into zones of extra-body concepts corresponds to a symbolic representation, which seems not what the brain uses. In terms of completeness, the theoretical results in Weng 2011 [91] established that the network-wise uniform DP of a generative DN (DGN) is sufficient to learn any complex Finite Automaton (FA), incrementally, immediately, and error-free, but grounded in the physical world, without using any handcrafted symbolic representations like the FA.

In the remainder of this section, we discuss the brain at four scales — from global to local — brain, cortex, layer, and neuron.

I. Brain Scale: Cortex Seems General Purpose

Regulated by the genome, the cerebral cortex develops a processing hierarchy [40] through extensive experience. Before we describe WWN at different scales of the hierarchy, we outline different scales of this hierarchy.

At the brain scale, the cortex is organized as pathways, as illustrated in Fig. 1. Each sensing modality (visual, auditory, touch, etc.) corresponds to different sensory pathways, which may not be single in the cortex as shown in Fig. 1. Each of these pathways occupies different cortical areas before converging to the frontal cortex where they are integrated and linked with the motor pathways.

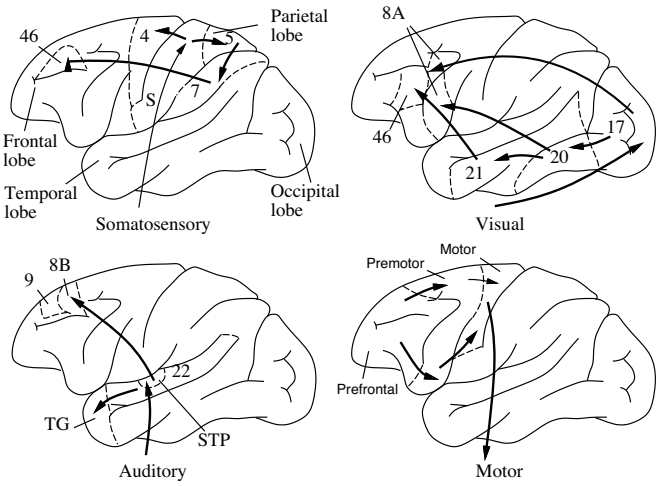


Fig. 1. The major lobes of the cerebral cortex and major cortical pathways. Upper left: somatosensory pathways. Upper right: visual pathways. Lower left: auditory pathways. Lower right: motor pathways. The somotosensory, visual and auditory pathways converge to the frontal cortex where the motor pathways drive motor neurons (effectors). Only bottom-up links are shown, but every one-way connection is in fact two-way realized by two one-way connections. The numbers marked are Brodmann codes for brain areas. Adapted from [40].

Our developmental model indicates that the motor area (i.e., it represents actions) is the main hub for multi-modal integration. For example, the feature neurons in the visual area for visual "car" stimuli and the feature neurons in the auditory area for auditory "car" stimuli should all link to the same action in the motor area by the Hebbian learning mechanism — the action of pronouncing "car" by the agent itself (mirror neurons), as illustrated in Fig. 5 later.

The correlation between visual features and the auditory features exists. Such correlations are represented by multi-modal feature areas in the brain but such areas are found to be relatively small. This seems due to the fact that such correlations are likely not as strong as feature-and-action correlations. Many sensory features lead to the same action but many sensory features do not often occur at the same time.

Therefore, each sensory pathway consists of a network of cortical areas, but they converge into motor areas. Neurons in early cortical areas have smaller receptive fields than those in later areas. But the neurons in the motor area has the largest receptive fields. The motor area appears the largest multi-modal area in the brain.

Computationally, the WWN model indicates that each pathway has the following characteristics.

1) Brain Processing from Two Signal Sources: Sensor and Effector: The brain faces a major challenge. It does not have the luxury of having a human teacher to implant symbols into it as its skull is closed. Thus, it must generate internal representations from the two signal sources: the sensors and the effectors (motors).

Thus, the brain can be modeled as a highly recurrent regressor $(\mathbf{z}', \mathbf{x}') = r(\mathbf{x}, \mathbf{z}, M(t))$ which maps sensory input-output $\mathbf{x}$ and motor input-output $\mathbf{z}$ using its current memory

$M(t)$ to the updated version $(\mathbf{z}', \mathbf{x}')$. The two-way connections to the effector $\mathbf{z}$ enables the brain to not only control the effector, but also to learn from the actions that are either generated externally or internally. The external generation takes place when, e.g., a child learns passively while the teacher directly guides his hand. The internal generation occurs when, e.g., the brain is practicing.

The output from the Lateral Geniculate Nucleus (LGN) to sensory source (retina) does not exist in primate, but this is not always so with other lower animals. The output to an early sensory area is useful for attention and reduction of noise and motion blur. Unlike some other emergent models discussed in Section V, we consider that the purpose of attention is to generate desired behaviors, not to reconstruct images. The lack of connection from LGN to retina in the primate central nervous system seems to support this perspective.

2) Top-down input corresponds to both top-down attention and context: The top-down input reflects the status of the motor (e.g., want cars or pedestrians) which enhances the response of the best-matched neurons. This information reflects both attention (in terms of enhancing response) and top-down context (in terms of immediate goal). Further, the duration of such a top-down context is variable, depending on how long the object in the sensory input has been present and attended.

3) Bottom-up input corresponds to both bottom-up saliency and feature match: The bottom-up input reflects the detected features that have survived competition. The competition includes both goodness of bottom-up match (as the inner product of input and weights) and the effect of top-down attention in the past. That is, they are reflected as bottom-up saliency this time — the more often attended in the past the more neurons are recruited for representing the same range of features and, thus, the better the bottom-up match.

4) Cross modality plasticity: Sur et al. [82] showed that the extensive sectioning (i.e., cutting) of the ascending auditory inputs from the inferior colliculus (from the ears) into the Medial Geniculate Nucleus (MGN) of the default auditory pathway of the newly born ferret causes retinal projections (visual sensory nerves) to innervate (i.e., axons grow into) the now "job less" MGN. This is equivalent to altering the source of bottom-up area (called X) of the auditory Y from the normal auditory source to the visual source. Interestingly, Sharma, Angelucci & Sur [74] showed that the rewired auditory cortex Y displayed visual orientation map similar to V1, but less orderly. This biological work indicates that a cortical area can emerge (i.e., develop) to work for very differently signal sources, normally auditory but now visual.

5) A brain area as general-purpose learning: The cross-modality plasticity is unlikely restricted only to the auditory pathway. Different cortical areas have shown other similar plasticity properties [83],[21, pp. 270-283]. When we talk about a subarea Y in the brain B, we use X, Y, and Z, where the bridge Y has its two banks X and Z. When we talk about brain B, we use S, B, and M, as a special case of X, Y, and Z. We predict that each brain area Y is a general purpose "bridge" that develops to predict the signals in both "banks" X and Z. By general purpose, we do not mean that there is no genetically modulated pre-disposition, such as the default neuronal resource. Yes, this "bridge" is two-way. Although one can call X to be bottom (sensory) and Z top (motor), this is not always necessary. For example, the Lateral Intraparietal Cortex (LIP) links the dorsal and ventral pathways for which it is not necessary to tell which is bottom and which is top. In our model, X and Z are largely treated symmetrically by Y.

J. Cortex Scale: Prescreening before Integration

Every cortical area has six laminar layers, regardless of its function. Layer L1 is the superficial layer and layer L6 is the deepest layer. Weng et al. 2008 [99] reasoned that L4 and L2/3 are two feature detection layers with L5 assisting L2/3 and L6 assisting L4, in the sense of enabling long-range lateral inhibition. Such long-range inhibitions enable different neurons to detect different features.

The DN model was informed by the work of Felleman & Van Essen [25], Callaway and coworkers [9], [105] and others (e.g., [33]). There are no top-down connections from L2/3 to L4, indicating that L4 uses unsupervised learning (U), competition among bottom-up components. L2/3 features are supervised (S) as the survived bottom-up features in X and survived top-down features in Z integrate in L2/3 to generate the bridge representation Y for $X \times Z$. Weng et al. 2008 [99] reported that such a *paired* hierarchy USUS led to better recognition rates than the unpaired SSSS alternative. An important function of such paired cortical layer is to prescreen bottom-up features to only allow top-match neurons to fire so that top-down input does not hallucinate very weak or absent bottom-up features.

To simplify our following discussion, we will omit the prescreening layers and simply model each cortex area as a single layer that directly integrates the bottom-up and top-down flows.

K. Layer Scale: Lobe Component Analysis

As discussed above, each cortical area has two feature layers, L4 and L2/3. The layer here means a feature layer, either L4 assisted by L5 or L2/3 assisted by L6.

1) Two conflicting criteria and working and long-term memory: Each cortical area faces two conflicting criteria that many neural networks face: fast adaptation of working memory and large stable long-term memory in the same cortical layer.

2) LCA: dual optimality: The above problem is resolved by distinguishing related memory from unrelated ones. The sparse coding principle (Olshaushen & Field 1996 [59]) allows only few neurons to fire. Those firing neurons correspond to best matched filters for the current neuronal input and are considered the working memory for the current input. Other neurons in the layer are long-term memory for the current input. Therefore, the role of working memory and long-term memory is dynamic. The inhibition in each feature layer is assisted by the corresponding assistant layer. This working memory and long-term memory model published in Weng & Luciw [97] is different from typical psychological explanations. Rather, our model is based on cortical anatomy.

Different from Olshaushen & Field 1996 [59] who considered sparse coding as a part of their objective function, we consider sparse coding as a result from neuro-anatomically observed lateral inhibitory connections, which suppress the firing of all irrelevant long term memory neurons for this cortical context.

We have developed a model called Candid Incremental Covariance-free (CCI) LCA [101], [97]). Mathematically, a lobe component is the first principal component of a region in the random input space $X \times Z$ that the neuron belongs to, assigned through neuronal competition. This model has a dual optimality: (1) Best representation for space — the smallest average error in the sense of the principal component. That is, the least average error for the cortex to represent the cortical input using a limited number of feature neurons. (2) Best estimation of representation through different learning times — the smallest average error from the starting time (the birth of DN) up to the current time, among all the possible estimators, under some regularity conditions.

L. Neuronal Scale: In-place Learning

Although the above dual optimality is described at the layer scale because of the neuronal interactions, the learning must be performed in-place by each neuron, without requiring any extra-cellular mechanisms. The above spatial optimality (1) gives the Hebbian increment direction. The above temporal optimality (2) gives the best step size.

The step size of each neuron depends on the individual firing age of the neuron. The CCI LCA [101] optimal step size scheduling requires that every neuron to have a self-stored age and age-dependent plasticity schedule. Thus, every neuron in a layer has a different automatically determined (optimal) learning rate.

III. Network Model

When you speak, do your muscle activities shape your brain's thinking? Due to the top-down connections from your motor areas, WWNs indicate that they do.

As we discussed earlier, we consider a general purpose area Y, which is connected with its sensory area X and its motor area Z, as illustrated in Fig. 2. The order of areas from low to high is X, Y, Z.

A. Area Function

During "prenatal" learning, the c neurons of A in $\{X, Y, Z\}$ need to initialize their synaptic vectors $V = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_c)$, and the firing ages $A = (a_1, a_2, ..., a_c)$. Each synaptic vector $\mathbf{v}_i$ is initialized using the input pair $\mathbf{p}_i = (\mathbf{b}_i, \mathbf{t}_i)$, consisting of the bottom up input $\mathbf{b}_i$ and the top-down input $\mathbf{t}_i$, $i = 1, 2, ..., c$, at the first c time instances. Each firing age a_i is initialized to be zero, $i = 1, 2, ..., c$.

After "birth," at each time instant, each area A computes its response $\mathbf{r}'$ from its input $\mathbf{p} = (\mathbf{b}, \mathbf{t})$ based on its adaptive part $N = (V, A)$ and its current response $\mathbf{r}$, regulated by the attention vector $\mathbf{b}_a$, and updates N to N':

$$(\mathbf{r}', N') = f(\mathbf{b}, \mathbf{r}, \mathbf{t}, \mathbf{r}_a, N)$$

where f is the area function described below. The attention supervision vector $\mathbf{r}_a$, having the same dimension as $\mathbf{r}$, is used to softly avoid the area A from excessively learning background. In our simulation, it suppresses all the A neurons to zeros except $3 \times 3 = 9$ ones centered at the correct object location. Biologically, the vector $\mathbf{r}_a$ is driven by other connected brain areas and is not very accurate during early ages, as a child does learn something from backgrounds. This need for the soft internal attention vector $\mathbf{r}_a$ is expected to be removed in future more powerful modeling of brain-like development.

The area A can be any of the areas X, Y, Z. The sensory area X and the motor area Z also compute and update in this unified way. But X does not have bottom-up input and Z does not have top-down input since they are nerve terminals. Receptors and muscles are nerve terminals.

B. Hextuple Fields of Each Neuron

It is known that each neuron in an early cortex has a (classical) *receptive field* (RF), corresponding to a field in the retina. The *effective field* (EF) of a neuron is the scope of its effect on the motor ends — the 3-D array of *muscle elements* over the body, called *"muxels"* here for short. We feel that the term "effective" is symmetrical to "receptive", and is different from the term "projective" used by Sejnowsky 2006 [72] to mean downstream targets. The term "projective" does not imply effectors. These two fields, RF and EF, corresponds to bottom-up flows — from pixels to muxels.

It has been documented that each cortical neuron receives three types of connections, bottom-up, top-down and lateral [25], [19]. Evidence [48], [67] suggests that multiple representations of an object exist, each specific to either perception (ventral) or action (dorsal). It is known biologically that feedback connections have been widely present in the dorsal and ventral streams [25], [13], [41]. They may contribute to illusory contours [33] (e.g., Kanizsa triangle). However, the operational roles of top-down connections are not clear [30] and neither are their representational effect.

To understand internal representation, we introduce that the receptive field (RF) of a neuron should contain three parts: *sensory RF* (SRF), *motor RF* (MRF) and *lateral RF* (LRF), respectively. By "lateral" we mean connections with neurons in the same area that are strongly correlated or anti-correlated, similar to Felleman & Van Essen's sensory and motor hierarchies [25]. The effective field (EF) of each neuron should also include three parts: *sensory EF* (SEF), *motor EF* (MEF), and *lateral EF* (LEF), respectively. Primate LGN does not project back to the retina, which is a special case, indicating that SEFs of later neurons have LGN cells as the highest possible resolution. See Fig. 2(a) for these six new fields — *hextuple fields*.

Two neurons are connected if they co-fire often (i.e., Hebbian learning). Therefore, for each neuron, three pairs are similar:[1] SRF with SEF, MRF with MEF, and the excitatory parts of LRF and LEF.

[1]But the corresponding weights are not the same, as the firing ages of the pre- and post-synaptic neurons are typically not the same.

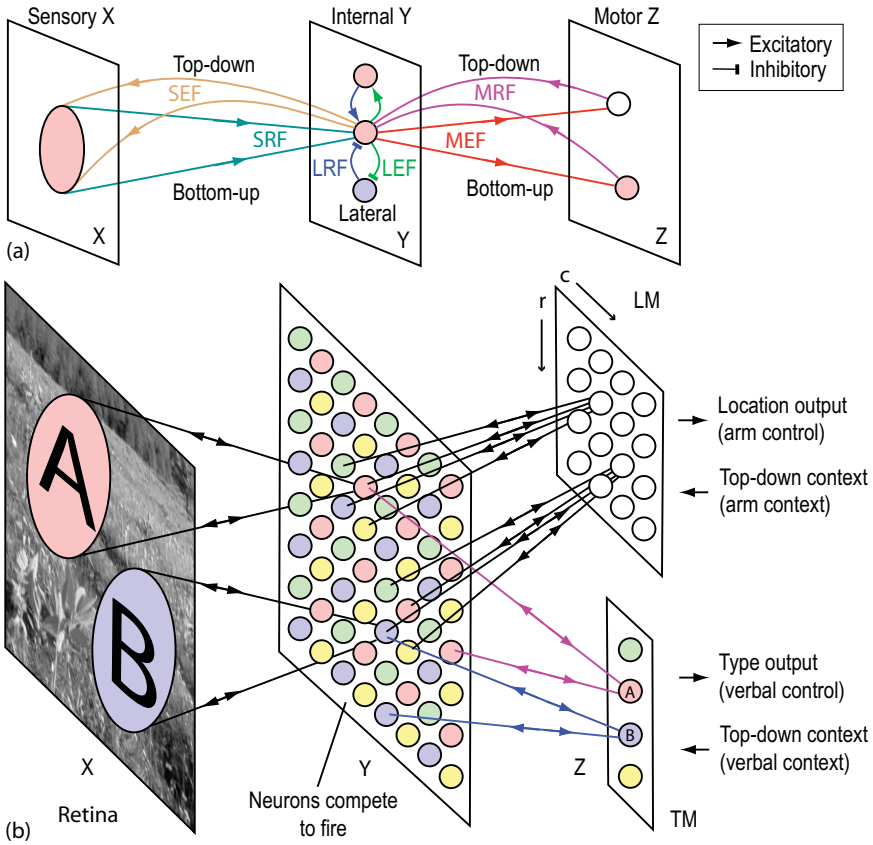


Fig. 2. A simple WWN and the network hextuple fields. (a) The hextuple fields for each neuron: SRF, MRF, LRF, SEF, MEF, and LEF, thus are highly recurrent. (b) As DN embodiment, a WWN has three areas: retina X, simple brain Y, and motor Z. The Z area has two concept areas LM and TM. Each wire connects if the pre-synaptic and post-synaptic neurons have co-fired. The weight is the frequency of pre-synaptic co-firing when the post-synaptic neuron fires. Within each cortical area, each neuron connects with highly correlated neurons using excitatory connections but connect with highly anti-correlated neurons using inhibitory connections. This forces neurons in the same area to detect different features in SRF and MRF. These developmental mechanisms result in the shown connections. Every Y neuron is *location-specific* and *type-specific*, corresponding to an object type (marked by its color) and to a location block (2×2 each). Each LM neuron is location-specific and type-invariant (more invariance, e.g., lighting-direction invariance, in more mature WWNs). Each TM neuron is type-specific and location-invariant (more invariance in more mature SWWs). Each motor neuron pulls all applicable cases from Y. It also top-down boosts all applicable cases in Y as top-down context. A two-way arrow means two one-way connections, whose two synapses are generally not the same. All the connections within the same area are omitted for clarity. All LM and TM neurons have global SEFs.

To facilitate discussion, we denote Y as a simple brain in Fig. 2(b). In general, Y should include neurons with small and large SRFs. Each Y neuron with a specific receptive field is responsible for detecting a feature at its specific location with the specific scale.

Not to forget motor areas, we trace the dorsal stream further to the *Location Motor* (LM) area and the *Type Motor* (TM) area. LM loosely represents the frontal eye field (FEF) and the arm reaching control area in the pre-motor and the motor cortices [13]. TM loosely represents the ventral frontal cortex (VFC) and the following verbal control area in the pre-motor and motor cortices [13]. Therefore, the muxels in LM and TM hubs are "meta" muxels, representing instances of abstract actions.[2] LM and TM roughly correspond to implicit skill and explicit knowledge, respectively, discussed in cognitive science. In our WWN example, we only teach LM to learn two concepts: vertical direction (row) and horizontal direction (column). Likewise, we only teach TM to learn one concept: type. According to our hextuple fields, motor neurons are not only action output ports, but also input ports for the network.

Fig. 2 gives an example of the resulting hextuple network representation throughout a WWN which consists of one retina buffer, an internal area (Y), and two motor areas (LM and TM). The amount, richness, and sophistication of its behaviors are limited by the resource available and its experience (e.g., "living" age).

Before discussing how WWN abstracts, we must first see how it computes.

C. Area Computation

For each neuron we need to refer to their input source neurons and output target neurons. Corresponding to SRF, SEF, MRF, MEF, LRF, and LEF of a neuron, we change receptive field (RF) to input neurons (IN) and change effective field (EF) to output neurons (ON) and thus define SIN, SON,

[2]The motor hierarchy in the motor area have neurons with larger MRFs similar to SRF for the sensory cortex, giving neurons representing "meta" muxels.

MIN, MON, LIN, and LON, respectively, as the *hextuple neuronal sets* of each neuron.

Thus, the SINs give the components of bottom-up $\mathbf{x} \in X$, the LINs the lateral $\mathbf{y} \in Y$ and the MINs the top-down $\mathbf{z} \in Z$, corresponding to its three parts of weights, $\mathbf{v}_x$, $\mathbf{v}_y$, and $\mathbf{v}_z$, respectively. Namely, each SIN corresponds to a component in the vector space X, etc..

Each cortical area uses lateral connections to enable its neurons to find their roles. There are two types of lateral connections, inhibitory and excitatory. On one hand, adaptive *excitatory* connections are extensive at early ages to generate a smooth map to globally cover a rough "terrain" but gradually become selective and local to fit the detail of the "terrain". Highly correlated cells form a clique and they are connected by excitatory connections. On the other hand, adaptive inhibitory connections find highly anti-correlated cells, and therefore they connect cells from different cliques. The major purpose [97] of inhibitory connections is to allow only few best responding neurons (winners) to fire and update, and other neurons (e.g., weakly responding backgrounds) do not fire to distract so that they can also keep their long-term memory intact.

In our LCA model of cortex, lateral inhibitions within area Y are simulated by the top-k competition mechanism, not by actual inhibitory connections, to quickly identify the top winners within each network update. This is especially useful when the software or hardware cannot run fast enough (e.g., update the entire network at 1kHz or above) to quickly sort out the winner in real time (every 30ms). In the experiment, we hand-picked the value of k, assuming that the value is largely gene pre-dispositioned. The adaptive lateral excitation [49] within area Y can encourage a smooth representation in early development and lateral prediction (e.g., edge filling in Kanizsa triangle).

As explained in the sparse coding theory [59], [97], it is important that only few top winner neurons in each area fire and update so that those that do not update serve as long term memory for the current context.

Consider an area A in $\{X, Y, Z\}$. In general, each neuron in A has a weight vector $\mathbf{v} = (\mathbf{v}_b, \mathbf{v}_t)$, corresponding to the area input $(\mathbf{b}, \mathbf{t})$, if both bottom-up part $\mathbf{b}$ and top-down part $\mathbf{t}$ are applicable to the area. Otherwise, the missing part of the two should be dropped from the notation. Its pre-action is the sum of two normalized inner products:

$$r(\mathbf{v}_b, \mathbf{b}, \mathbf{v}_t, \mathbf{t}) = \dot{\mathbf{v}} \cdot \dot{\mathbf{p}}, \tag{1}$$

where $\dot{\mathbf{v}}$ is the unit vector of the normalized synaptic vector $\mathbf{v} = (\dot{\mathbf{v}}_b, \dot{\mathbf{v}}_t)$, and $\dot{\mathbf{p}}$ is a unit vector of the normalized input vector $\mathbf{p} = (\dot{\mathbf{b}}, \dot{\mathbf{t}})$.[3] The inner product measures the degree of match between these two directions $\dot{\mathbf{v}}$ and $\dot{\mathbf{p}}$, because $r(\mathbf{v}_b, \mathbf{b}, \mathbf{v}_t, \mathbf{t}) = \cos(\theta)$ where θ is the angle between two unit vectors $\dot{\mathbf{v}}$ and $\dot{\mathbf{p}}$.

[3]In our more recent experiments with much large backgrounds than the ones reported here, we found that it is beneficial to first subtract the means of $\mathbf{x}$ and $\mathbf{z}$ from each so that the average "brightness" in $\mathbf{x}$ and $\mathbf{z}$ does not affect the inner product. To avoid a zero denominator for a constant vector, each subtracted mean is reduced by the standard deviation of the digital quantization noise.

Consider the area A to be Y, connecting with bottom-up input area X and top-down input area Z. The area Y with many neurons has a set of cluster centers:

$$\{(\mathbf{v}_x, \mathbf{v}_z) \mid \mathbf{v}_x \in X, \mathbf{v}_z \in Z\}. \tag{2}$$

Each center $(\mathbf{v}_x, \mathbf{v}_z)$ is the center of the corresponding Voronoi tile in the area's input space $X \times Z$. The competition of neurons discussed below means that all the samples in the tile is represented (quantized) by the center. Each center $(\mathbf{v}_x, \mathbf{v}_z)$ is an instance of co-occurrence which humans generally have no language term (symbol) to correctly identify. Thus, each center is not pure linguistically in any human language — can not be precisely described as the concepts of the external environment in any natural language. This linguistic impurity should be true for all internal neurons inside the brain skull — those neurons that cannot be directly supervised by the external environment.

It is known in electrical engineering that positive feedbacks may cause uncontrollable oscillations and system instability. Lateral reciprocal inhibitory connections require many fast iterations with unpredictable oscillations — an unsolved great challenge faced by the nonlinear control theory and many-cell neurodynamics. Our computational theory for a cortical area, the Lobe Component Analysis (LCA) [97], uses a top-k mechanism — a highly nonlinear mechanism — to explain that lateral inhibitions enable neurons in each area Y to sort out top winners within each time step t_n, $n = 1, 2, 3, ...$ Let the weight vector of neuron i be $\mathbf{v}_i = (\mathbf{v}_{bi}, \mathbf{v}_{ti})$, $j = 1, 2, ..., c$. For simplicity, considering $k = 1$, the single winner neuron j is identified by:

$$j = \arg \max_{1 \le i \le c} r(\mathbf{v}_{bi}, \mathbf{b}, \mathbf{v}_{ti}, \mathbf{t}).$$

Suppose c is sufficiently large and the set of c synaptic vectors distributes well (mathematically, the density of the c points well approximates the observed probability density in the parallel input space. Then, with a high probability the winner (nearest neighbor) neuron j has its both parts match well:

$$\mathbf{v}_{bj} \approx \mathbf{b} \text{ and } \mathbf{v}_{tj} \approx \mathbf{t}$$

not counting the lengths of these vectors because of the length (contrast) normalization in $r(\mathbf{v}_b, \mathbf{b}, \mathbf{v}_t, \mathbf{t})$.

Consider area A to be area Y. We would like to have the response value y_j to approximate the probability for $(\mathbf{x}, \mathbf{z})$ to have $\mathbf{v}_j = (\mathbf{v}_{xj}, \mathbf{v}_{zj})$ as the nearest neighbor. For $k = 1$, only the single winner fires with response value $y_j = 1$ and all other neurons in the area do not fire $y_i = 0$ for $i \neq j$.

In general $k > 1$, a dynamic scaling function dynamically shifts and scales each pre-action potential r_i so that the top-1 winner has a response value $y = 1$ and the $(k + 1)$-th and weaker neurons respond with zeros, avoiding the undesirable effect of any static threshold. The above represents c update equations with c unknowns in $\mathbf{y}$. Without a need to explicitly

solve c simultaneous equations, each dynamic function[4] g depends on values in $\mathbf{y}$ at time t_n to give the updated response y_i but for the next time t_{n+1}.

As further discussed below, $\mathbf{z}$ may be supervised using abstract concepts, such as location and type. However, a few rounds of internal updates using the above c computations with c neurons quickly blur the linguistic boundaries in the top-down $\mathbf{z}$ with the concrete bottom-up $\mathbf{x}$. This is a computational account about the absence of any handcrafted boundary ("walls") that separate symbolic meanings even if $\mathbf{z}$ is supervised to represent abstract concepts. That is what we mean by "the brain's internal representation seems not pure linguistically in any human language."

D. Receptive Fields Are Selective and Dynamic

Conventionally, a receptive field has been computationally modeled as a more-or-less static field for a sensory neuron (e.g., detecting a feature in the field). The new hextuple concept means that a receptive field is *attention-selective* and *temporally dynamic* — a different subpart is active at a different time [27], depending on top-down attention and the competition in early areas.

RF is conventionally for a sensory neuron, not a motor neuron. However, the SRF of the "pink" motor neuron (e.g., "person" as type-A) in Fig. 2(b) is a union of the overlapping SRFs of all "pink" Y neurons (e.g., as instances of "person" at different retinal locations). Thus, the SRF of each motor neuron is global, but selective and dynamic, since only a few Y neurons win to fire at any time.

This dynamic, selective SRF explains why each TM neuron is *locationally invariant* and *type specific*. Similarly, every LM neuron is *type invariant* and *location specific*.

Likewise, an MRF is also selective and dynamic, e.g., different motor actions boost a V1 neuron at different contexts. An MRF is typically disconnected (e.g., each Y neuron connects one neuron in LM and TM, respectively).

In general, we argue that the brain represents all human communicable concepts through its motor areas as "hubs". Any human communicable concepts can be produced by muscle contractions: Verbal concepts can be communicated through muscle languages — written, verbal, sign languages, etc.. Non-verbal concepts can be produces through muscle procedures — reaching, grasping and manipulation. Therefore, the motor vector $\mathbf{z}$ can be taught to represent any concept.

E. Concepts

Weng et al. [100] argued that the genome (developmental) program is body-specific (e.g., sensor-specific and effector specific) but not task-specific. In principle, any language and any concept about the external world can be learned after the programming, subject to available resources. The DP is sensor-specific and body-specific, in the sense that its X area is tied to a particular sensor (camera in this case) and its Z area is tied to a particular type of body effectors (e.g., each motor neuron corresponds to a finger). However, the DP is not task-specific. Of course, each particular WWN regulated by the DP is task-specific and environment-specific, because each WWN can be trained by a particular sequences of task in a particular social environment. In our experiment, the members of the first society S_1 partition all the fingers into two groups, LM and TM. The S_1 society uses canonical representation for TM and the row-column representation for LM, as we explained below. However, this is just a particular language in society S_1. Another society S_2 may use a different language, not necessarily a canonical one.

Let $C = \{(c_1, c_2, ..., c_n) \mid c_i \in \mathbf{Z}, i = 1, 2, ..., n\}$ consists of vector of n learned concepts, where each concept variable c_i has m_i possible values from $\mathbf{Z}$ (denoting the set of all integers), represented by m_i neurons in sub-motor area i, $1 \leq i \leq n$. In Fig. 2(b), the WWN has learned $n = 2$ concepts, location and type. The location concept has $m_1 = 4 \times 4 = 16$ possible values, and the type concept has $m_2 = 1 \times 4 = 4$ possible values.

In each sub-area, the neuron that has the highest response indicates the output from the sub-area. But the relative contrast among the neurons for the same concept indicates confidence. All the response values from every neuron is normalized to $[0, 1]$. Consider Fig. 2(b). c_2 represents the 2nd concept type, then $c_2 = j$ means that the type is of class j. Thus, the j-th neuron in TM have a value 1 and all other neurons in TM are zeros. This type of representation of concept is called *canonical representation* — each neuron represents a concept value.

Each motor sub-area typically specifies a value at any time, but this is not always necessary. When all the neurons in a motor sub-area are all nearly zero, this representation means "do not know" as motor output and "do not care" as motor input. During motor-imposed teaching, we set the corresponding neuron to 1 and all the other motor neurons in the same sub-area to be zero to indicate "absolutely sure".

The concept representation in WWN is emergent, not symbolic. Suppose that every concept variable has the same v possible values. Each combination of all c concepts corresponds to a motor state. The total number of symbolic states is v^c, exponential in c. The possible concepts can be object name (type), row, column, distance, scale, material, weight, color, viewing angle, lighting, relations between objects, etc. For $v = 4$ and $c = 22$, $v^c = 4^{20} = 16^{11} > 10^{11}$, larger than the number of neurons in the human brain. In contrast, the distributed representation in the motor area Z requires only vc neurons to represent all possible v^c symbolic states. For $v = 4$ and $c = 22$, $vc = 88$, instead of 16^{11}. Using numerical parameters such as synaptic weights, an emergent representation can interpolate among an exponentially number of unobserved vectors. This is a great advantage of an emergent representation over a symbolic representation.

[4]Suppose $r_1 \geq r_2, ..., \geq r_{k+1}$ and $r_{k+1} \geq r_i$ for all $i > k + 1$. Then $y_i = g(r_i)$, $i = 1, 2, ..., c$, where the function $g(r)$ depends on the ranked values r_1 and r_{k+1}: $g(r_i) = (r_i - r_{k+1})/(r_1 - r_{k+1})$, if $i \leq k$. It subtracts r_{k+1} from the input r and divides the difference by $r_1 - r_{k+1}$ so that the response vector $\mathbf{y}$ of the area has k positive components with the maximum value reaching 1. All remaining $c - k$ components in $\mathbf{y}$ are zeros: $g(r_i) = 0$ for all $i > k$.

F. Learning

The hextuple network representation in Fig. 2 should not be statically handcrafted for three major reasons [21], [63], [100]. First, new objects appear through a lifetime. Second, a network needs previously learned skills to autonomously learn more sophisticated skills. Third, handcrafted representations are suboptimal.

All the WWN weights learn through an incremental, interactive "seeing" and often supervised "acting" process. By supervised acting, we mean that the external environment (e.g., teacher) supervises in real time the motor port, which is used as both input and output by the WWN. The network grabs one at a time input pair $\mathbf{p} = (\mathbf{b}, \mathbf{l}, \mathbf{t})$ consisting of bottom-up input $\mathbf{b}$, lateral input from the same area $\mathbf{l}$ (excitatory only) and top-down input $\mathbf{t}$ to update the area A before the next input $\mathbf{p}$ is grabbed.

Consider area A to be Y. If only partial or none of Z are available, the network's self-generated values are used (i.e., practice during semi-supervision). Simulated pulvinar signals (early attention) allow only Y neurons in the 3×3 region centered at the correct location to fire and update during training. The exact neuronal location is unknown even during training, since at each location there are fewer neurons than all possible foreground objects and each neuron must report for multiple similar locations. Thus, the SRF of every Y neuron is contaminated by "leaked-in" background pixels, requiring the top-down representational effect discussed below.

The traditional error back-propagation models [103], [44] do not consider long-term memory, not suited for development as earlier skills must serve as long term memory for acquisition of later, more sophisticated skills. The Lobe Component Analysis (LCA) [97] not only has a long-term model, but also cast long-term and short-term memory in a dually optimal framework. LCA was compared with some well known methods in [97]. The learning in each area uses the same LCA procedure. LCA is similar to Self-Organization Map (SOM) [43] and LISSOM [53] but it optimally distributes the limited number of neurons of each area optimally in the input space $X \times Z$ — optimal Hebbian learning, spatially and temporally, as illustrated in Fig. 3.

Consider a general area A. The *spatial optimality* of A sets up the best target. With a limited number of neurons in each area, the set of all synaptic vectors is V. The best representation for each areal input $\mathbf{p} = (\mathbf{b}, \mathbf{t})$ is $\hat{\mathbf{p}}(V)$, whose error is $\|\hat{\mathbf{p}}(V) - \mathbf{p}\|$. Note that the quality of the representation $\hat{\mathbf{p}}$, as explained in [97], is measured as the error of reconstruction from the best matched synaptic vector in V but the actual reconstruction is not performed as the synaptic weight vector can not be read. The spatial optimality [97] identifies the theoretically best set V^* that minimizes the expected representation error: $V^* = \arg\min_V E\|\hat{\mathbf{p}}(V) - \mathbf{p}\|$.

The *temporal optimality* of A does the best for $V(t)$ at every time t through lifetime, by minimizing its expected distance to the best but unknown target $E\|V(t) - V^*\|$. Suppose that the neuron j with synaptic vector $\mathbf{v}_j$ is the top winner. This temporal optimality [97] leads to not only Hebbian direction $y\mathbf{p}$ but also the best step size $w(n_j)$, best in terms of the temporal optimality, for every update:

$$\mathbf{v}_j \leftarrow (1 - w(n_j))\mathbf{v}_j + w(n_j)(y\mathbf{p}) \tag{3}$$

where $w(n_j)$ and $1 - w(n_j)$ are the optimal learning rate and retention rate, respectively, both depending[5] on the firing age n_j of neuron j. See Weng & Luciw [97] for derivation. The real-valued firing age is updated as $n_j \leftarrow n_j + y$.

For example, a child is staring at a novel car (indicated by pattern A in Fig. 2) and his pulvinar suppresses other background sensing neurons as he attends. This leads to the firing of pink Y neuron in Fig. 2 that best matches the "car" image patch at the correct retina location. At the same time, his mother repeats "car, car," which excites, through child's the auditory stream, the child's motor neurons for pronouncing "car". (This association should have established before since when the child motor pronounced "car", his auditory stream heard his own "car" — co-firing.) This corresponds to the firing between the Y neuron and the pink motor neuron in TM in Fig. 2. Their synapse (both-way) is connected with the Hebbian increment yp_i where p_i is each active Y neuron. The learning of LM is analogous.

Let us carefully examine the case of perfect supervision of all the motor neurons. As the car sweeps across the retina against the complex background, the "car" motor neuron in TM is supervised to fire, and the correct location neuron is also supervised to fire. The pulvinar attention signal guarantees that Y neurons whose receptive field is far from the correct car location are suppressed. So, only Y neurons whose receptive field roughly coincide with the car region potentially can fire. However, the Y neuron that actually fires is the one whose weight vector matches the car best. First, consider the synapse from each firing Y neuron responding to the car at the *correct location* and the supervised-to-fire "car" motor neuron in TM. The Y neuron is the pre-synaptic neuron and the Z neuron is the post-synaptic neuron. The Hebbian learning uses this co-firing event to strengthen the synaptic weight (e.g., from zero to a non-zero weight). Thus, as the car sweeps across, the "car" neuron in TM automatically connects all the Y neurons that respond to "car" at different locations. This "car" motor neuron is then "car" specific but location invariant. Similarly, all other type neurons in TM are also type specific and location invariant, as illustrated in Fig. 2. The same principle applies also to every LM neuron. Each location neuron adds connections from Y neurons responding to different types but at the specific location. This enables every LM neuron to be location specific and type invariant, as illustrated in Fig. 2.

Because of the statistical average nature of the optimal Hebbian learning in Eq. (3), precise timing of action is helpful (e.g., the eyes attend to the current position of the hand), but is not always necessary. This is because an object will stay for a while. Imposition of an action for a new object

[5]The plasticity schedule $w(n_j)$ is probably genetically scheduled. For WWN, $w(n_j) = 1/n_j$ gives the optimal step sizes $w(n_j)$ for a fixed distribution of $y\mathbf{p}$ under some regularity conditions. For a practical, slowly changing distribution of $y\mathbf{p}$ due to network's "grow-up", $w(n_j)$ slowly increases from $1/n_j$ to $3/n_j$ in the "critical window" of "child " age. $w(n_j) = 3/N$ when n_j has reached a mature age, e.g., $n_j \geq N = 2000$, to maintain an small "adult" plasticity.

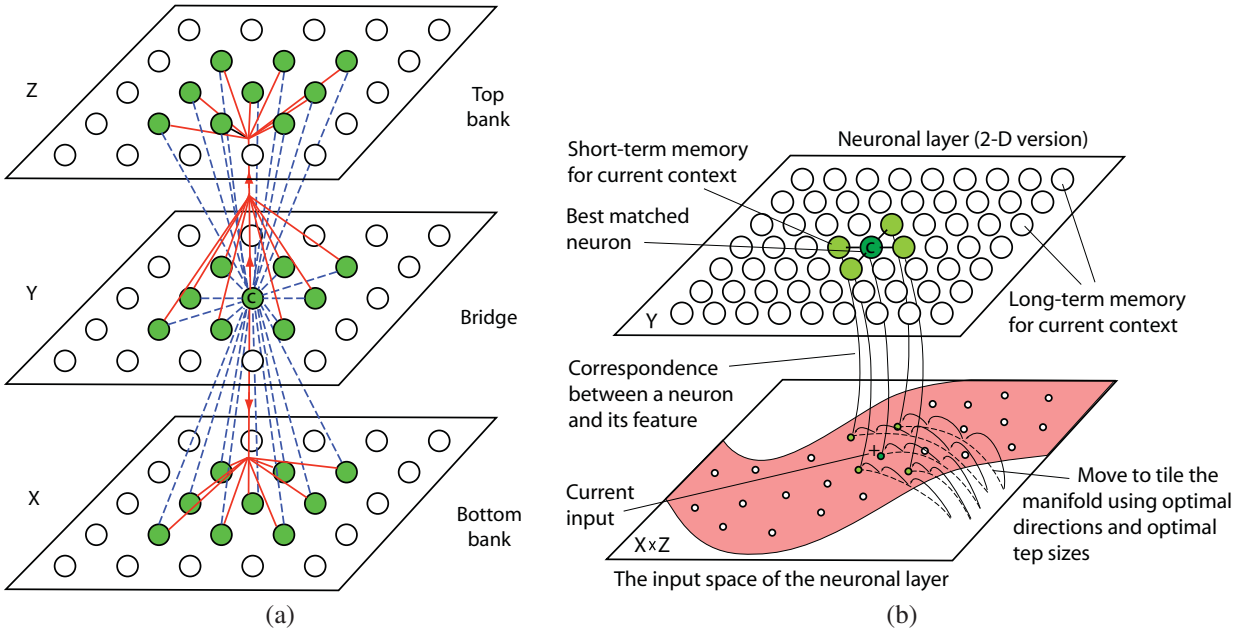


Fig. 3. Illustration of Lobe Component Analysis for bottom and top input spaces. (a) The default connection pattern of each neuron in area Y. The Y area is a bridge representation that serves its two banks: the bottom space X and the top space Z. All the connections are local but two-way. Blue: neuronal input; red: neuronal output. In the same layer, near neurons are connected by excitatory connections (for representation smoothness) and far neurons are connected by inhibitory connections (competition resulting in detection of different features by different neurons) (b) The meaning of the dual optimality of LCA. The upper area indicates the positions for the neurons in the 3-D Y area (i.e., several stacked layers within Y). The firing neurons (green) are context-dependent working memory for this context and those do not fire are context dependent long-term memory for this context. The lower area indicates the very high dimensional input space $P = X \times Z$ of the cortical area Y, but illustrated in 2-D. Each curve links the neuron in Y plane with its synaptic weight vector illustrated as the position in $P = X \times Z$. For simplicity, the weight vectors of Y represented in P as small dots define a Voronoi diagram in P. The magenta area in P indicates the manifold of the distribution of input data, which is typically very sparse in P and of a much lower dimension than the apparent dimension of P. The spatial optimality of LCA means that the target tiling by the Voronoi diagram in the pink area is optimal to minimize the representation error for $P = X \times Z$. The temporal optimality of LCA means that the neuronal weight of firing neurons must move toward their unknown best target the quickest through out the developmental experience. As illustrated, the updating trajectory of every neuron is a highly nonlinear trajectory. The statistical efficiency theory for neuronal weight update (amnesic average) results in the nearly minimum error in each age-dependent update, meaning that not only the direction of each update is nearly optimal (Hebbian direction), but also every step length (fully automatically determined).

does not need to always be aligned with the onset of the relevant visual information. A few second after the onset of the object, an action is supervised (or answer provided) which enables the agent to rehearse the action internally (i.e., speak to itself softly) while it continuously reinforces the object-action association through its Hebbian learning in Eq. (3).

On the other hand, it is incorrect to think that inter-stimuli interval is unimportant. For example, the inter-stimuli interval between a tone and an air-puff needs to fall within the range from 325ms to 550ms for the classical conditioning to be learned effectively by an animal [32], [79].

Suppose $k = 1$ in top-k competition in the Y area. Each Y neuron fires at 1. Then, it can be shown [91] that the above learning expression incrementally updates the synapse as the sample probability for the pre-synaptic neuron to fire conditioned on that the post-synaptic neuron fires.[6]

All "loser" neurons are not updated and their ages do not advance, serving as the long term memory relative to this context $\mathbf{p}$. Therefore, the role of each neuron as working-memory or long-term memory is relative and dynamic. If it fires, it is part of the current working memory and updates. Otherwise, it is part of the long term memory. Therefore, forgetting occurs only in the details of the nearest matched memory for "unconscious" refinement of skills.

[6]That is, the WWN can be considered a spiking network, simulating the Spike Timing Dependent Plasticity.

G. DN Algorithm

The following mechanic algorithm incrementally solves the task-nonspecific, optimal learning problem of the highly non-linear, highly recurrent DN, with WWN has an embodiment.

Algorithm 1 (DN): A DN has three areas, X, Y and Z. Y is always hidden. X is exposed to the external environment as it connects with sensors. Likewise, Z is exposed as it is connected with effectors. The internal brain area Y predicts the responses in *two banks* X and Z as the *bridge*.

1) At time $t = 0$, for each area A in $\{X, Y, Z\}$, initialize its adaptive part $N = (V, G)$ and the response vector $\mathbf{r}$, where V is the synaptic weights and G the neuronal ages.
2) At time $t = 1, 2, ...$, for each area A in $\{X, Y, Z\}$, do the following two steps repeatedly forever:

a) Every area A computes using area function f.

$$(\mathbf{r}', N') = f(\mathbf{b}, \mathbf{t}, N) \quad (4)$$

where f is the unified area function; $\mathbf{b}$ and $\mathbf{t}$ are area's bottom and top inputs from current network response $\mathbf{r}$, respectively; and $\mathbf{r}'$ is its new response.

b) For each area A in $\{X, Y, Z\}$, A replaces: $N \leftarrow N'$ and $\mathbf{r} \leftarrow \mathbf{r}'$.

If X is the sensor port S, $\mathbf{x} \in X$ is always supervised by the external environment. If Z is the effector port M, $\mathbf{z} \in Z$ is supervised only when the teacher chooses too. If the teacher does not supervise $\mathbf{z}$, the DN self-supervises (practices). Regardless $\mathbf{z}$ is supervised or not, $\mathbf{z}$ gives motor output if Z is an effector port.

Instead of considering Y as the entire internal brain, we can also consider Y as a subarea of the internal brain, where Y can be the spinal cord, the hindbrain, the midbrain, the forebrain, or a Brodmann area. The two banks X and Z compute themselves, just like the bridge Y, while they are partially supervised by other brain areas in addition to Y.

H. Concept Purity

The sensory space X (e.g., an image patch) and the motor space Z (e.g., a particular action) are both concrete. The brain has only limited choice (e.g., attention) for the sensory space X since it consists of projections from the natural world. The brain has more choices for the motor space Z since evolution has selected effectors that are useful for the agent. We propose that brain attended *concepts* are expressed through the motor area Z, since the other exposed end X is for sensors only.

Suppose that the area Y has enough neurons to well sample the sensory space X to a sufficient density that is sufficient to distinguish the concept values for the concept space C that the network will end up learning (e.g., location and type). Other environmental variations that the concept C does not learn need to be sufficiently sampled (e.g., use neurons with different default sizes of receptive field to detect different sizes of foreground object). In other words, for any $\mathbf{x} \in X$, every best matched Y neuron is *concept-pure*. A neuron (or an area) being concept-pure is different from an area having a crisp concept boundary, because the latter at least implies that some neurons in the region is concept-pure. A neuron that is not concept-pure means when it fires, one of multiple concepts can be true.

Mathematically, a neuron j is concept-pure is defined as follows. Let $P_j \subset P = X \times Z$ to be the subset of contexts so that the Y neuron j is the winner:

$$P_j = \{(\mathbf{x}, \mathbf{z}) \mid j = \arg\max_{1 \le i \le c} r(\mathbf{v}_{xi}, \mathbf{x}, \mathbf{v}_{zi}, \mathbf{z}), (\mathbf{x}, \mathbf{z}) \in P\}.$$

If all the contexts $(\mathbf{x}, \mathbf{z}) \in P_j$ has a single $\mathbf{z} \in C$, then we say that P_j is concept pure. That is, whenever $(\mathbf{x}_1, \mathbf{z}_1) \in P_j$ and $(\mathbf{x}_2, \mathbf{z}_2) \in P_j$, we have $\mathbf{z}_1 = \mathbf{z}_2$ Specifically, we may also say that a Y neuron to be location pure (c_1) or type pure (c_2). Note that each input $(\mathbf{x}, \mathbf{z})$ has two components. When $\mathbf{z}$ is a zero vector, the context $\mathbf{p}$ is a pure bottom-up case (free-viewing mode). Otherwise, it is a $\mathbf{z}$ top-down context and $\mathbf{x}$ bottom-up context integration case (goal-directed mode).

The concept-pure case is useful for understanding how the system works, although the exact concept-purity is not guaranteed with a limited-size network.

Theorem 1 (perfect motor output): Suppose that all Y neurons that have ever fired are concept-pure and the supervision at motor area Z is all correct. Suppose also that the network uses top-1 firing rule for Y and the society language uses a language-specific sub-area in motor Z for each concept. Then, every winner concept neuron in Z is correct for the particular concept value of c_i it represents and is concept invariant for all the other $n-1$ concepts learned by the network.

Proof: Since every Y neuron is pure in the combination of n concept (e.g., only one pair of location-type), it is impossible for the Hebbian learning to link an impure Y neuron with a Z neuron. All the Y neurons collected by each motor neuron through Hebbian learning are different cases for this particular concept value (e.g., a particular type, but for different locations). Thus, each motor neuron is pure for its concept value (e.g., the type value is "car") and invariant to all other $n-1$ concepts (e.g., location). Mathematically, all Y neurons form a perfectly fine partition of the space P according to Z concept labels (values), and every Z neuron perfectly collects all its cases from Y. Then, all Z neurons form a perfect super (i.e., coarse) partition of P. ■

The requirement of all concept pure Y neurons may need unrealistically larger number of Y neurons. Top-k competition with $k > 1$ but relatively small can perform interpolation with sparsely populated Y neurons in the lower dimensional manifold inside the high-dimensional P, as reported in [98], [52]. Recently, Weng 2011 [91] reported that there is a type of generative WWNs that incrementally learn one time frame at a time like the WWNs here but guarantee to give 0% error for all training data observed so far and are optimal, in the sense of maximum likelihood, for all disjoint testing data.

The WWN does not treat features in Y as a "bag-of-features" used in some other scene classification methods [24], [73] to reduce the number of training samples, because of the inner-product-based neuronal response for Z. The location of each element in a vector $\mathbf{x}$ affect the outcome of the inner product.

I. A WWN Example

As discussed earlier, a developmental network is not meant for a specific task but instead for incrementally learning skills required for performing a variety of tasks suitable for its age and useful for its target applications. Our current emphasis of experimental studies is on the properties, power and limitation of the theory, method and algorithm. In Section IV, we will discuss a variety of networks. Here let us see a WWN example so that we can continue our theoretical discussion using this example. More detail of this WWN example will be presented in Section IV

We trained a WWN using images like those in Fig. 4(a). The object contour is approximated by an octagon by default. The refinement of object contour needs synapse maintenance, discussed elsewhere in [88], which automatically cuts off synapses that have bad matches since their pre-synaptic input

is from backgrounds. Synapse maintenance was not used in the experiments reported here.

To simulate a shortage of neuronal resource relative to the input variability, we used a small network, five object image patches of a single scale, and many different natural backgrounds. Both the training and testing sets used the same 5 object image patches, but different background images. As there are only 3 Y neurons at each location but 5 objects, the WWN is 2/5 = 40% short of resource to memorize all the foreground objects. The Y area optimally uses the limited resource by implicitly balancing the trade off between type concept representation and location concept representation. Therefore, each Y neuron must deal with various misalignment between an object and its receptive field, simulating a more realistic resource situation.

We are now ready to see what a trained WWN can do. We will see that WWN does not just do pattern recognition. As outlined in Table I, it does not require the human programmer to model each concept but instead enables un-modeled concepts (e.g., location and type here) to be learned interactively and incrementally as actions; it enables such concepts to serve as goals (supervised or self-generated) but in general serve as attended spatiotemporal equivalent top-down contexts; and it enables such goals to direct perception, recognition and behavior emergence.

J. Free-viewing Mode

Without top-down inputs from motor areas, the network operates in the free-viewing mode. This mode is also called *bottom-up* attention [36], [37] — a salient learned object "pops up" from backgrounds. Within WWN the saliency is learned, supported by the neuro-anatomic studies on cortical connections from which we derived our computational model, although a human new born has a set of inborn behaviors. For example, a toy is salient to a young boy but not salient to an elderly. The shapes of leaves on a tree is salient to a plant biologist but not so to a young boy. Our prediction that bottom-up saliency is largely learned is also consistent with experience-dependent saliency reported by Lee et al. [46].

As reported in Fig. 4(b), the network gave respectable performance after only the first round (epoch) of practice. After 5 epochs of practice, the network reached an average location error around 1.0 pixels and a correct classification rate over 99%. This is the first solution to the joint attention-recognition problem in unknown complex backgrounds with a practical-grade performance in free-viewing mode. The dynamic selective SRF of all motor neurons are essential for the success.

This new capability is potentially applicable to a wide variety of vision problems that currently do not have a general-purpose technical solution, such as autonomously finding objects (e.g., people, cars, man-made structures) from complex backgrounds, and while driving a car autonomously controlling a pan-tilt head through LM signals according to the location of the object of interest found.

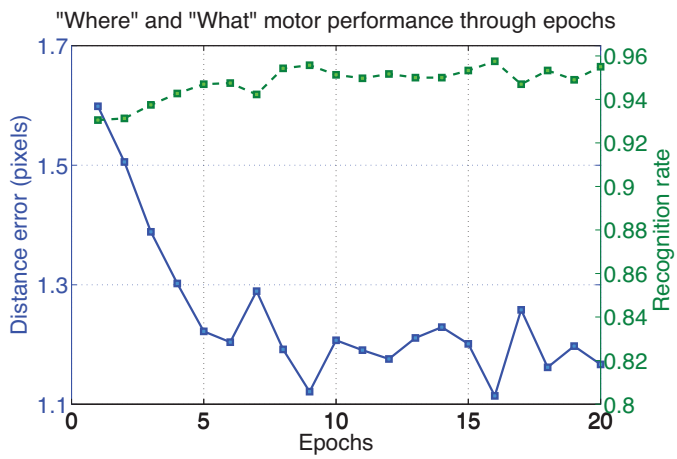
Fig. 4(c) will be discussed later. Fig. 4(d) shows the Y class map from the disjoint testing in the free-viewing mode, which shows that most neurons are almost class-pure, except a few around the decision boundaries. This is because each Y neuron fires and learns only when a foreground object is present, top Y winners report excellent matches of a single type. The top-down representational effect discussed below further discounts leaked-in background pixels (due to limited neurons), since the co-firing wiring enables the correct motor neuron to send the correct top-down signal to the correct Y neuron during training and practice. The LCA optimality [97] contributed to the superior purity of V2 neurons under a limited number of neurons and experience. Fig. 5(e) gives two examples of outputs in the free-viewing mode.

This brain-inspired object representation scheme is different from the proposed appearance-kept shift-circuits proposed by Anderson & Van Essen 1987 [3], implemented by Olshausen, Anderson & Van Essen 1993 [58], and extended by Tsotsos et al. 1995 [86] for an internal master feature map originally schematically proposed by Treisman 1980 [85]. The WWN model does not require the existence of such a holistically object-aware, topography-kept, scale- and location-invariant master map. The first problem with such a master map is the absence of the mechanism that is necessary for the autonomous development of such a holistic master map from experience. The second problem is that it does not reconcile the distributed nature of brain's internal representations emerging from autonomous self-organization using sensory inputs and motor inputs. The third problem is that this master map requires two separate circuits — one for identification of the location and the scale of attended "spot light"; the other for the normalization from this "spot light" into the assumed master map. The fist circuits seems harder, without a general-purpose computational solution so far. The brain-inspired WWN scheme here not only copes with the functions of the above two circuits using the same set of mechanisms, a uniform developmental program of the WWN can develop the entire brain-like circuit.

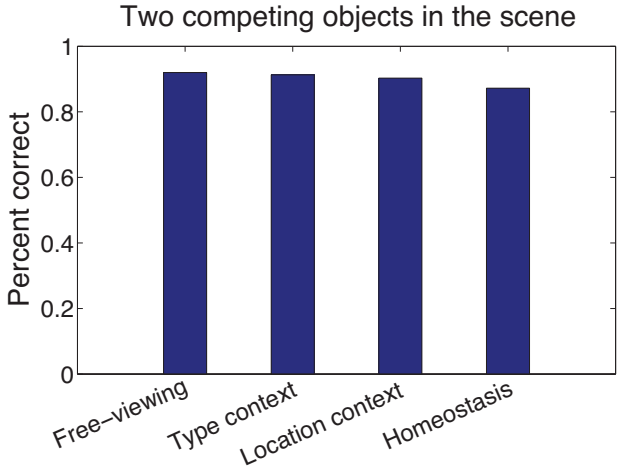
From the above idea that each brain area serves as a bridge to predict its two banks, the most basic bridge between the sensory port and the motor port is the "skeleton" base. This "skeleton" base for the somatosensory motor system is the spinal cord which is developed earlier in life. The "skeleton" base for visuomotor system seems to be the thalamus, which contains the LGN and the pulvinar. The pulvinar bi-directionally connects with all the areas in the forebrain [58]. The areas in the higher brain, which develop slightly later than the "skeleton" bases but also co-develop with the bases, add "flesh" to the "skeleton" bases to refine the corresponding prediction between the sensory port and the motor port. Namely, brain's internal representation can be regarded as adding more and more areas between any pair of source area X and target area Z.

The deep learning scheme proposed by Lee & Mumford 2003 [45], Hinton et al. 2006 [34] considers the brain circuits as a deep cascade of areas. We have experimentally shown that a cascade of four areas X-Y1-Y2-Z performed worse than one which adds direct connections between Y1 and Z [50], and also performed worse than three areas X-Y-Z using less resource [88]. Adding more "flesh" areas to the

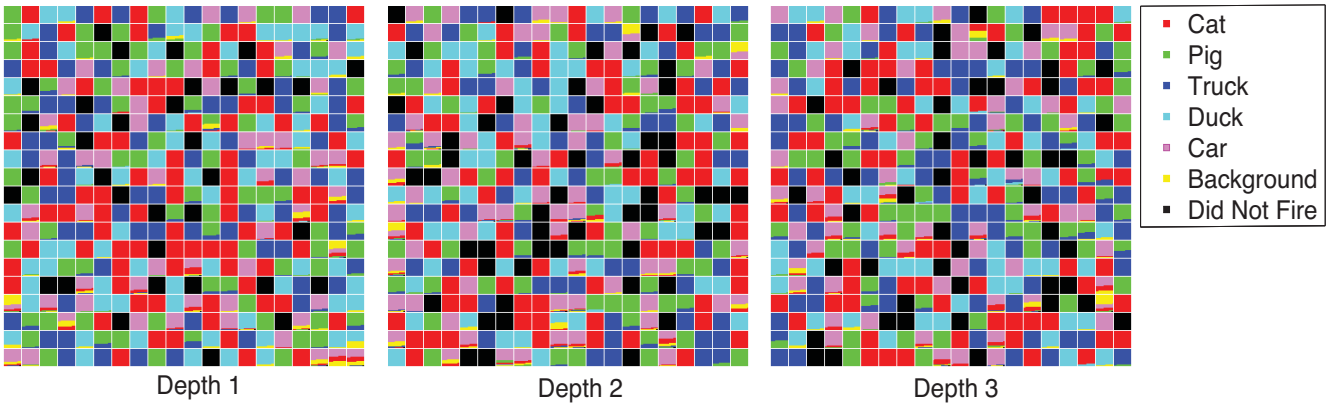
(a)



(b)



(c)



(d)

Fig. 4. The performance of a limited-size WWN. (a) Sample image inputs. Five object patches "car", "table", "penguin", "person", "horse" pasted at any location on a randomly selected natural background image. Later versions of WWN used arbitrary contours for the foreground objects since the model allows any contours. (b) The average errors of the reflexive actions, reaching and telling the type (recognition) during free-viewing in unknown complex natural backgrounds, which improve through epochs of learning experience. (c) Performance when input contains two learned objects: reflexive (free-viewing), two types of goal-directed recognition (top-down type-context and location-context), and fully autonomous goal-switching (homeostasis). (d) Type-concept representation map (for TM) of Y, using top-1 winning rule, disjoint test, in the free-viewing mode. It has an array of cells, each representing a Y neuron (20 rows, 20 columns, depths 1 to 3 corresponding to the thickness of a cortical layer). In each cell, the area of a color is proportional to the corresponding probability of the type. If all the neurons are type-pure, all the cells have a single color. As Y has a limited number of neurons, these Y neurons are not all type-concept pure.

"skeleton" bases is beneficial for improving the prediction between intermediate representations, but they should help the skeleton base, instead of acting alone. The X-Y-Z three-area architecture is functionally sufficient (see Theorem 1 in Weng 2011 [91]). The dual optimality of LCA [97] further justifies that each area predicts the best based on its limited resource and limited amount of experience.

K. Goal Directed Perception

It is known [17], [13], [41] that visual *top-down* attention as operational bias has two types, *location* based and *object* (or feature) based. As we discussed above, the top-down signal from a motor action can represent any human communicable concepts, the goal-directed recognition scheme below is applicable to general abstract concepts as goals.

Fig 5(b-d) illustrates the 3-stage process as a complete link

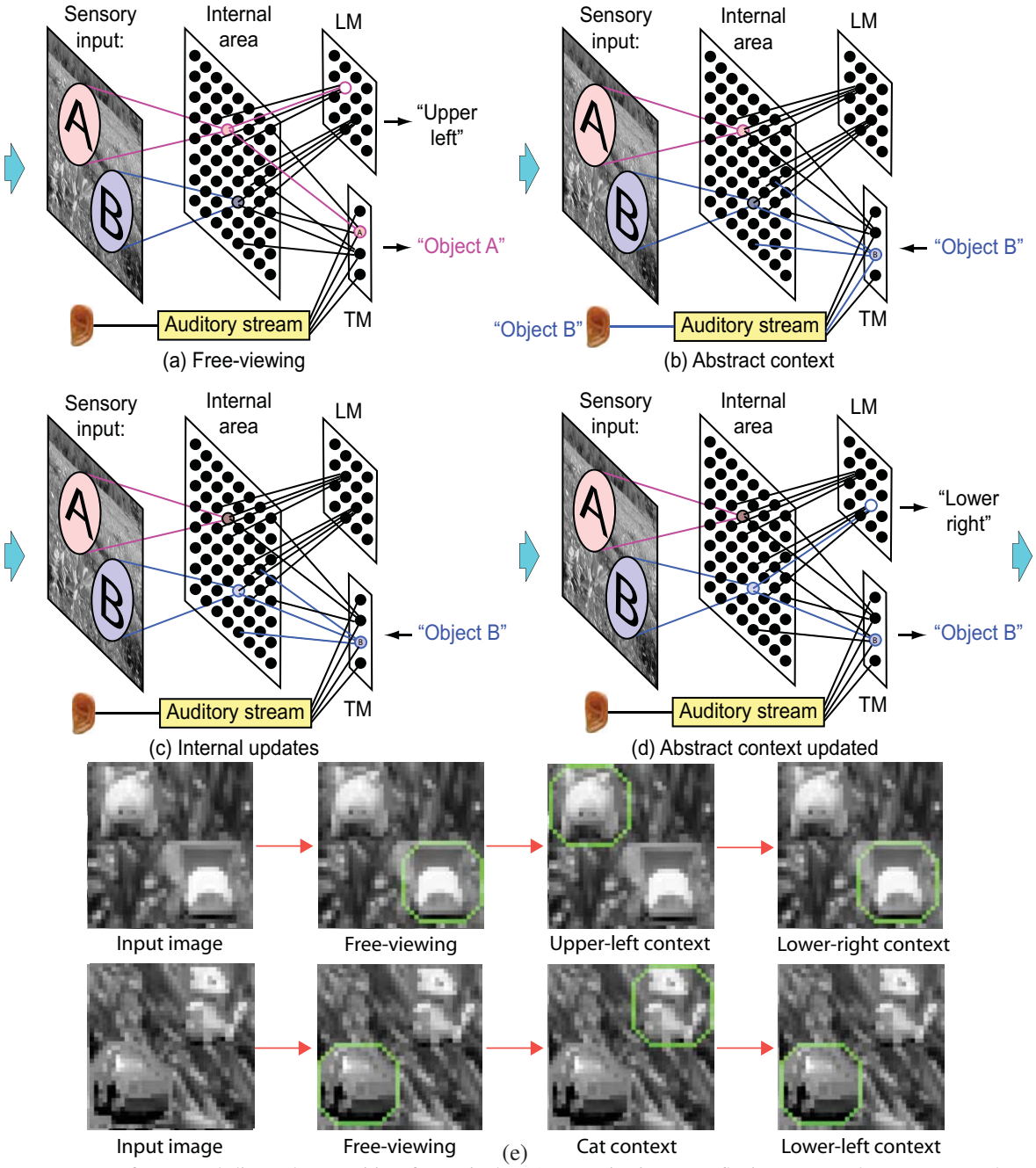


Fig. 5. How WWN performs goal-directed recognition from pixels. (a) Free-viewing — reflexive, no top-down context. The WWN now attends object A and it is at "upper left". (b-d) With top-down context — deliberative. (b) Abstract concept. A friend stated an abstract concept "Object B." Through the auditory stream, WWN gets "Object B" firing in its TM area as top-down context. (c) Internal updates. Winners among top-down boosted Y neurons now fire (one here), with all pixels taking into account. (d) Abstract concept emerged. The firing Y neuron sends the response to LM and LM, where one reporting the abstract location and the other confirms the abstract type. (e) Some examples of goal-directed recognition by a trained WWN. "Context" means top-down context. A green square indicates the location and type action outputs.

in a series of general purpose WWN goal-directed perception:

...
$\Rightarrow$ Abstract concept in Z: action and/or goal arise
$\Rightarrow$ Internal updates in Y: goal-directed competition
$\Rightarrow$ Abstract concept in Z: action and/or goal arise
$\Rightarrow$ Internal updates in Y: goal-directed competition
...

As we explained earlier, the term "goal" is a special case of top-down spatiotemporal context. This process involves not only goal-directed perception, but also goal-directed recognition and goal-directed reasoning in the sense of from abstract concept to another abstract concept. Goal-directed reasoning while examining the bottom-up evidence is also called deliberative reasoning.

The first stage is *"abstract concept in Z"*. The top-down context is an instance of motor action now, representing a value of a query concept (location or type). It can be either self-generated, externally injected (motor-end supervision or sensory-end communication), or a mix of both. Fig. 5(b) shows that the top-down concept is communicated via the ears (e.g., from a teacher). Recall why visual and auditory pathways share the same action as we discussed earlier. The concept(s) represented by the motor here is general-purpose, as it can be any other human communicable concept (e.g., goal or criteria). The firing TM neuron(s) sends boosting signals to all its SONs in Y, using the Hextuple representation from TM (SEF). As a special case of this stage is the top-down attention [17], [13], [41] — location-based, type-based and more, via motor hubs.

The second stage is *"internal updates in Y"* — computation with (abstract) top-down context and (concrete) bottom-up pixels (foreground and background) using the entire network's Hextuple representations (see Fig. 5(c)). All the above SONs in Y are boosted, increasing their chance to win. The "originating" motor neurons together with the boosted and now firing Y neurons conceptually correspond to what is called "motor imagery" [60] during which a human mentally simulates a given action. Further repeated neuronal computation for all neurons in Y, LM, and TM using their SINs, MINs and LINs results in what conceptually called the "mental imagery" by Shepard & Metzler's [75] where the top-down context corresponds to an imaginary rotation action.

The third stage is *"abstract concepts in Z"*. The Y winners send signals to MONs (e.g., now involving all related motor areas) using the entire network's Hextuple representations. The motor areas (LM and TM) display the result of goal-directed recognition as an instance of the emergent concepts and action (see Fig. 5(d)) but it can represent an instance of any abstract concept(s) in general. Koch & coworkers 2009 [64] reported cells that have such multimodal invariance in the hippocampus and the entorhinal cortex of the macaque monkey, as those areas are multimodal like TM and LM in Fig. 5. The WWN model gives a computational explanation for the emergence of such multimodal invariant cells. The hippocampus and the entorhinal cortex has access to effectors for, e.g., navigation.

This autonomous process goes on and on for an open-ended long time. We propose that this process characterizes the most basic mode of brain thinking, as argued by Weng 2011 [90]. The concepts that the network thinks about are emergent, this thinking process is applicable to a wide variety of practical concepts that can emerge. This thinking capability is rooted in experienced associations through the emergent *internal* Hextuple representation, instead of an *externally* handcrafted symbolic representation. As we discussed earlier, a statistically consistent sensorimotor association implies a meaning from the physical world. Namely, this is "meanings-from-physics". In particular, it is not based on mathematical logic. A fuller discussion as in Weng 2010 and 2011 [90], [91] about the completeness of this thinking-and-reasoning process needs the tool of automata and is beyond the spatial scope of this work.

We tested the above learned WWN for goal-directed recognition with two competing objects in each retina image, at four possible quadrants to avoid overlapping. As shown in Fig. 4(c), the success rates are 96% from the type context to reason location and 90% from the location context to reason type.

To allow the network to self-generate its own top-down contexts (i.e., abstract "thoughts") like an autonomously "living" animal, we use its *homeostatic mode*. The currently two firing motor neurons in LM and TM get suppressed (simulating temporal depletion of synaptic vesicles which package neural transmitters) and relatively other neurons are boosted concurrently (simulating the disappearance of lateral suppression from the previous winner). WWN correctly recognized the "runner-up" object (in LM and TM) under this *homeostatic mode* with an average success rate 83% (see Fig. 4(c)).

L. Top-down Representational Effect

To understand this new subject, we need to see firstly how neurons are distributed in the input space and then how top-down signals sensitize relevant bottom-up components.

First, the earlier expression for neuronal learning can be rewritten as $\mathbf{v}_j \leftarrow \mathbf{v}_j + w(n_j)(y\mathbf{p} - \mathbf{v}_j)$. Thus, the amount of vector change $w(n_j)(y\mathbf{p} - \mathbf{v}_j)$ is proportional to the vector difference $y\mathbf{p} - \mathbf{v}_j = \mathbf{p} - \mathbf{v}_j$ when $y = 1$. We call it *the distance-sensitive property*. With this property, we have the *square-like tiling* theorem:

Theorem 2 (Square-like tiling): Suppose that the learning rule in a self-organization scheme has the distance-sensitive property. Then the neurons in the area move toward a uniformly distribution (tiling) in the space of areal input $\mathbf{p}$ if its probability density is uniform.

Proof: We give a geometric proof. See Fig. 6. Suppose that the density in S is uniform, as otherwise the same proof holds if local area is weighted by probability density. If an input $\mathbf{p}$ fall into the pink region in Fig. 6(a), the top winner neuron is its nearest neighbor. The Voronoi region of a point is the region in which every point has the point as the nearest neighbor. Each segment of the Voronoi region is the equal-distance border between the two points (neuronal weight vectors). An elongated Voronoi region means that statistical pullings are not isotropic. As explained in the goal-directed recognition in Fig. 6, the statistical pulling from observed samples move the points toward a distribution in

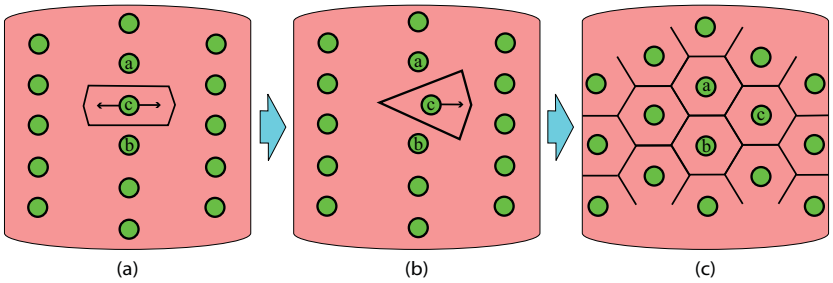


Fig. 6. The square-like tiling property of the self-organization in a cortical area. In a uniform input space, neurons in an layer self-organize until their Voronoi regions are nearly isotropic (square-like to nearly hexagons in 2-D) . (a) The Voronoi region of neuron c is very anisotropic — elongated horizontally — resulting horizontal pulling is statistically stronger. (b) A horizontal perturbation leads to continued expected pulling in the same direction (rightward in this case). (c) Through many updates, the Voronoi regions are nearly isotropic, ideally regular hexagons but generally square-like.

which all Voronoi regions are all isotropic. With a finite number of points, exact isotropy is impossible. In 2-D, squares are sufficiently isotropic and hexagons are slightly more so. In n-D, we call such nearly stable tiling square-like tiling. ■

Note that "move toward" does not state how fast. The speed of self-organization depends on the optimality of the step sizes. The temporal optimality of LCA deals with the speed.

Second, as shown in Fig. 7, learning using top-down inputs sensitizes neurons to action-relevant bottom-up input components (e.g., foreground pixels) and desensitize to irrelevant components (e.g., leaked-in background pixels). This is true during operation, when top-down input is unavailable during free-viewing. This is called the *top-down representational effect*. We have the top-down effect theorem:

Theorem 3 (top-down effect): Given a fixed number of neurons in a self-organization scheme that satisfies the distance sensitivity property, adding top-down input from motor Z in addition to bottom-up input X enables the quantization errors for action-relevant subspace X_r to be smaller than the action-irrelevant subspace X_i, where $X = X_r \times X_i$.

Proof: Given input $\mathbf{x} = (\mathbf{x}_r, \mathbf{x}_i) \in X$ where $\mathbf{x}_r \in X_r$ and $\mathbf{x}_i \in X_i$, suppose the nearest matched vector is $\mathbf{v} = (\mathbf{v}_r, \mathbf{v}_i)$. The quantization error $\mathbf{e} = \mathbf{x} - \mathbf{v} = (\mathbf{x}_r - \mathbf{v}_r, \mathbf{x}_i - \mathbf{v}_i) = (\mathbf{e}_r, \mathbf{e}_i)$, where $\mathbf{e}_r = \mathbf{x}_r - \mathbf{v}_r$ and $\mathbf{e}_i = \mathbf{x}_i - \mathbf{v}_i$. Consider $x \in R = [a - \delta/2, a + \delta/2]$ where the quantizer a is at the middle of R and x uniformly distributed in R, the quantization error is $e = x - a$. For uniform distribution of x in R, the standard deviation of its quantization error is $\rho = \sqrt{Ee^2} = \delta/\sqrt{12}$. For each component $\rho_r = \sqrt{e_r^2} = \delta_r/\sqrt{12}$ and $\rho_i = \sqrt{e_i^2} = \delta_i/\sqrt{12}$, where δ_r and δ_i are indicated in Fig. 7(a). A square-like tiling gives $\delta_r = \delta_i$. Therefore $\rho_r = \rho_i$ with top-down input. Next, consider using the top-down input $\mathbf{z} \in Z$ in learning and define $\mathbf{p}' = (\mathbf{x}, \mathbf{z})$. As $\mathbf{z}$ is independent of $\mathbf{x}_i$, we have the nonlinear map of $\mathbf{z} = f(\mathbf{x}_r, \mathbf{x}_i)$ in Fig. 7(b). Let $\partial \mathbf{z}/\partial \mathbf{x}_r = \mathbf{a}$, we have $\delta'_z = \mathbf{a}\delta'_r$. As $\mathbf{z}$ is not constant $\|\mathbf{a}\| > 0$. As illustrated in Fig. 7(c), one side of a square tile has a squared length $\delta_r'^2 + \delta_z'^2$ and the other side $\delta_i'^2$. The square tiling property gives $\delta_r'^2 + \delta_z'^2 = \delta_i'^2$. As $\delta'_z = \|\mathbf{a}\|\delta'_r$, we have $(1 + \|\mathbf{a}\|)^2\delta_r'^2 = \delta_i'^2$. Thus, $\delta'_r/\delta'_i = 1/(1 + \|\mathbf{a}\|) < 1 = \delta_r/\delta_i$. The larger the action relevance $\|\mathbf{a}\|$, the larger the sensitization. ■

This theorem gives two consequences (due to $\delta'_i > \delta'_r$): First, action-relevant bottom-up inputs are salient (e.g., toys and other Gestalt effects). Thus, we need to reconsider the conventional thinking that bottom-up saliency is static and probably totally innate. Second, relatively higher variation through a synapse gives information for cellular synaptic pruning in all neurons, to delete their links to irrelevant components.

M. Y Sub-areas

The internal area Y may contain multiple subareas. However, the forebrain is not a cascade of Brodmann areas, but a network of many Brodmann areas.

An earlier area (e.g., V2) links with not only the next area (e.g., V3) and previous area (e.g., V1), but also other later areas (e.g., inferior temporal area IT, medial temporal area MT, and the frontal cortex) and other earlier areas (e.g., LGN). This is a connection pattern universally found in the visuomotor pathways as the large tables in the survey of Felleman & Van Essen [25]. This complex connection pattern of the vision system has been puzzling in terms of computational reasons. However, the "bridge-and-banks" theory here explains that any two areas that are statistically correlated considerably should be connected. Therefore, the brain shows the complex connection pattern drawn by Felleman & Van Essen [25]. In particular, the brain is not a cascade of areas. *Any two neurons in the brain that are significantly correlated should be connected for better prediction for all "banks".*

For example, a motor sub-area that reports a feature with a small receptive field (e.g., a short edge) needs to be directly connected bi-directionally with a sensory area where the scale of the receptive fields are roughly correct (e.g., LGN and V1). A motor sub-area that reports an object type (e.g., car) needs to be directly connected bi-directionally with all sensory areas whose receptive fields may contain an object of the type (e.g., V1, V2, V4, and IT). The WWN that deals with multiple object scales is called WWN-5 reported in [78].

IV. Experimental Examples

In this section, we discuss how a general purpose DP model here can be used to develop a variety of networks each having different skills. In the future, a single, large, general-purpose network could learn potentially all such skills. It could

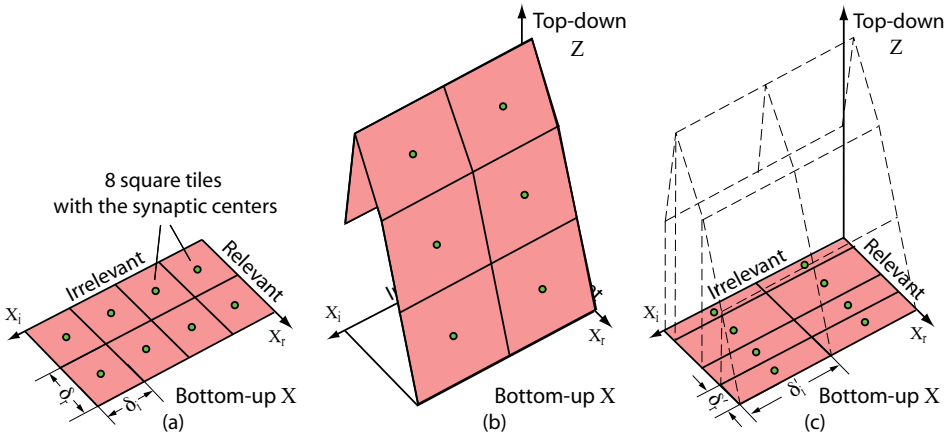


Fig. 7. The top-down representational effect — top-down inputs sensitize the response for relevant bottom components although which are relevant is unknown. (a) Without top-down input, square Voronoi tiles in the bottom-up space give the same quantization width for irrelevant component X_i and the relevant component X_r: $\delta_i = \delta_r$. All samples in each tile is quantized as the point (synaptic vector) at the center. (b) With top-down inputs during learning, square tiles cover the observed "pink" manifolds, indicating the local relationships between X_r and Z. (c) When top-down Z is not available during free-viewing, each tile is narrower along direction X_r than along X_i: $\delta'_r < \delta'_i$, meaning that the average quantization error for relevant X_r is smaller than that for irrelevant X_i.

continue to learn more complex mental skills through scaffolding — early simpler skills enable autonomous learning for later more complex skills, with or without human interactive supervision. This is largely a theoretical paper. For more detail of each experiments, the reader is suggested to read the paper cited.

A. Multiple Objects in Complex Background

This seems the first time using a general-purpose vision system, without assuming a hand-crafted concept about the environment (e.g., object, as the concept of object seems not present at the birth, per Jean Piaget [61]). This has been made possible by three tightly intertwined novelties: (1) un-modeled concept (taught at motor instead handcrafted into the system), (2) concepts as goals (emergent at the motor end as top-down goals for perception), and (3) goal-directed perception (both bottom-up $\mathbf{x}$ and top-down $\mathbf{z}$ are integrated in the internal cognitive matching), as summarized in Table I.

Each input retina image has a foreground image patch of scale 19×19 pasted into a randomly selected background image of scale 38×38. Thus, 75% pixels were from unknown backgrounds when a single foreground image is used. The sizes of the network areas are as follows. Retina area: 38×38; Y: 20×20; LM: 20×20; TM: 5×1, one for each of the 5 types of object. These numbers correspond to the required motor resolution. PP and IT areas, with the same size as the original Y, were inserted between the original Y and LM and between the original Y and TM, respectively. A slightly better performance was observed without PP and IT, better than Fig. 4(b-c). However, PP and IT are probably useful for actions with coupled motor neurons — multiple motor neurons fire concurrently. All default inter-area connections are global, two-way, except that only the bottom-up connections exist between retina and Y, which is 19×19, the scale of the objects trained and tested. (Variable object scales have been dealt with in WWN-5 [78], where the Y area has sub-areas each of which has a different scale for SRFs.) All default connections, serving as early over-connections, are pruned automatically during learning, leading to sparse connections as shown in Fig. 2.

The training images and testing images used different 38×38 background images randomly extracted from large images available at http://www.cis.hut.fi/projects/ica/imageica/. The training set consists of composite images, each consisting of one of the 5 objects (image patches) pasted at one of 20×20 locations of randomly extracted backgrounds images. This amounts to $5 \times 20 \times 20 = 2000$ images. Going through these 2000 images corresponds to one epoch. The training is fully incremental and supervised. Given each pair of input image and supervised motor response, the network updates three times, before the next pair is fed. The test set used the same foreground object image patches but different randomly extracted natural backgrounds.

For the Y area, bottom-up pre-screening is necessary for practical performance. From each input image, there are many moderate bottom-up matches (e.g., a car detector can get many moderate matches from a bush background). Consequently, a top bottom-up foreground match (e.g., a person match) can lose to an incorrect bottom-up background match (e.g., car match with bush) simply because the top-down bias boosts the latter (e.g., car bias). This is called *top-down hallucination*. There is also a similar need for pre-screening top-down matches if many motors neurons fire concurrently. Pre-screening is not a phenomenon of using rigid rules. Instead, it seems the best "guess" for the statistics of bottom-up inputs and top-down inputs so that biological structures are developed to facilitate the quickest estimation of statistics (i.e., using prior distributions in ancestor generations). Interestingly, the cerebral cortex appears to develop highly adaptive 6-layer architecture [25], [9], [19] to facilitate the quickest estimation of statistics, called pre-screening here, as explained below.

The above *m*atch-and-*c*ompetition mechanisms realize a

function: $\mathbf{y} = f_{m,c}(\mathbf{x})$ with input $\mathbf{x}$ and output $\mathbf{y}$. The subscripts m, c denote *m*atch and *c*ompetition layers, respectively, corresponding to specific laminar layers in the cortex. The cortex has 6 layers, L1 to L6. L1 has mostly axons to serve as information transmission "highway". The bottom-up input is matched in L4, with its competition through lateral inhibition assisted by L6. Thus, the above $f_{m,c}$ is replaced by L4 for m and L6 for c to generate the *pre-screened bottom-up* input: $\mathbf{y}_4 = f_{L4,L6}(\mathbf{x})$, meaning that the bottom-up input $\mathbf{x}$ is pre-screened at L4 with matching in L4 and lateral inhibition generated through L6 [25], [9], [19]. Likewise, the top-down input $\mathbf{z}$ is pre-screened in L2 whose lateral inhibition is assisted by L5, to generate *pre-screened top-down* input: $\mathbf{y}_2 = f_{L2,L5}(\mathbf{z})$. The integration of $\mathbf{y}_4$ and $\mathbf{y}_2$ is a neuron-to-neuron integration with a narrow input field (e.g., 1×1) in L2/3. For notational simplicity, we denote the bottom-top integration layer as L3: $\mathbf{y} = f_{L3,L5}(\mathbf{y}_2, \mathbf{y}_4)$. Recently, highly narrow radial cones have been reported in L2/3 [107], consistent with such small size of $f_{L3,L5}$ input fields. Note that the above functions $f_{L4,L6}$, $f_{L2,L5}$ and $f_{L3,L5}$ use the same match-and-competition mechanism as $f_{m,c}$ above.

In the WWN trained and tested, L4 and L2 have a thickness of 3 — having three feature detectors at each pixel location. However, there are 5 possible objects to be detected at each pixel location. This situation of limited resource requires a trade-off between type specificity and location specificity among all the L4 neurons. This trade-off is realized optimally through LCA.

The goal-directed recognition tests in Fig. 4(c) used disjoint backgrounds and the same foreground object image patches. The test data set consists of all combinations of two different objects at two different quadrants.

To test the performance for dealing with larger variability of each object class, we conducted another experiment. We used 25 3-D objects such as different toy cars and animals. Each is placed on a rotary table against a gray background so that the object is roughly centered, simulating the situation where overt attention (e.g., camera pan and tilt) has already brought the object to the center of the field of view. The DP developed WWN to recognize 3-D objects viewed from any of the 360° viewing angles. The Y area has a single layer without prescreening, as the background is not an issue in this experiment. The number of Y neurons (20×20) is limited: If each Y neuron is considered a quantizer of all the viewing angles, on average 90° of object viewing angle variation has only four Y neurons to quantize. With such a limited network size, the network still reached a classification rate of 96.9% (without using temporal context). When the size of Y area was increased to 30×30 and 40×40, the 3-D recognition rate reached 99.2% and 99.67%, respectively.

The remainder of this section serves as a summary of other related experimental results to additionally support the richness of the brain-inspired spatial processing presented here. This paper is not the primary publication venue for the experimental results below. Citations are provided for readers who are interested in more detail.

As the theory predicted, the shape of the foreground object contour does not have to be square. For example, WWN-2 [38] WWN-3 [51] used an octagon as the default shape of the bottom-up receptive field. Hebbian learning further weaken connections that do not correlate with the post-synaptic firing because it is from the background. WWN-2 [38] has dealt with a single object in complex backgrounds and WWN-3 [51] has coped with multiple objects in complex backgrounds. WWN-4 reported that direct connections between earlier sensory areas and motor areas gave better results (i.e., shallow connections, instead of a cascade of areas). WWN-5 [78] handles different object scales. To more accurately remove background pixels from each neuron that detects an arbitrarily shaped object, synapse maintenance has been incorporated into the WWN [88].

In addition to type-based object detection, location-based object recognition, and free-viewing object attention [51], the DN has also been experimentally tested on stereo [77] where the motor output corresponds to stereo disparities, natural language processing using inputs from the Wall Street Journal [102], and "thinking-like" transfer learning [56].

The richness of the tasks that a DN can perform depends on the sensors, the effectors, the computational resources, and the learning experience.

B. 25 3-D Objects with 360° of Viewing Variation

In Luciw & Weng [52], this theory was tested for the MSU-25 Data Set — 25 3-D real objects taken from any of the 360° viewing angles. Each object was roughly centered in the image, on a rotary base. Thus, only TM is used but not LM. The image frames not used for training was used for testing (disjoint test). A limited size network after 5 epochs of training reached a 96% recognition rate if each image was fed into the network individually — one shot recognition. However, if the test images were treated as video through time and the motor output is allowed to have multiple firing neurons to keep the confidence of past recognition, the recognition rate further increased to about 100%. This is an example of goal-directed perceptual reasoning, using the confidence of past recognition as a dynamically changing goal emerging from the TM area.

C. Outdoor Vehicles and Non-vehicles

This theory was also tested by Luciw & Weng [52] for a data set from the video sequences taken from a driving car in various driving and lighting conditions (e.g., under an overpass). It contains 225 vehicle images and 225 non-vehicle images, each sized 32×32. The network was trained to distinguish vehicles from non-vehicles. When 25% of the data was used for training, the network reached around 95% correct recognition rate. When the images are treated as video and the network uses the last motor output as "goals", the network performed "goal-directed perception" and reached a recognition rate of above 99%.

D. Stereo Perception without Explicit Stereo Matching

The theory here has also been tested by Solgi & Weng [77] in a very different setting — stereo without performing explicit stereo matching. We believe that without explicit

stereo matching is the way the brain deals with binocular vision. Each input to the network consists of two image rows of 20-pixel long each, one from the simulated left eye and one from the right, taken randomly from many natural images. The left and right image (rows) are relatively shifted by a stereo disparity from -8 to 8, which means that both two images have unknown non-overlapping part, depending on the underlying stereo disparity. In this experiment, LM is used but not TM. LM is trained to produce the underlying stereo disparities. All the tests are disjoint, from different source images. Without using top-down "goal", the network reached an average error around 1.6 pixels. Using motor top-down "goal" enabled the network to reach a sub-pixel average accuracy — about 0.6 pixels.

E. Natural Languages with Each Word as an Image

If the network is fed with an image at a time, where the image corresponds to a snapshot of a word, the network can be taught to produce the context state (e.g., parallel meanings as concepts) at the motor end, in a sense similar to the state of a Finite Automaton (FA). However, FA cannot generalize. The theory here has been tested for natural language processing using corpora from the Wall Street Journal as reported in Weng et al. 2009 [102] and early language acquisition and generalization (e.g., subclass-to-class generalization, subclass-to-subclass generalization) as reported in Miyan & Weng 2010 [56].

V. Prior Models

Existing models fall into two large categories, symbolic and emergent. A symbolic model requires the human programmer to provide a set of symbolic concepts about the extra-body environment, e.g., objects to be dealt with, a model of every object (e.g., parts and the relations among parts). In contrast, an emergent model does not require such a set of symbolic concepts. It uses numeric rules to enable internal representations to emerge from the sensory sources and the motor sources. Since no static symbolic set is sufficient for an open array of unknown tasks, it seems that only an emergent model is potentially task-nonspecific.

Weng 2012 [92] argued that the term "connectionist" is misleading and does not well characterize the more restrictive emergent property of brain-like representations and suggested the term "emergent representation" for brain-like representations.

A. Symbolic

Many computer vision methods [20], [87], [1], [106], and robotic methods that simulate child learning (e.g., see the symbolic methods reviewed in [6]) use symbolic representations, based on a monolithic 3-D model or a monolithic 2-D appearance model. Much of the intelligence of such models is from the human programmer, instead of the machine, since it is the human who understands each machine task and translates his understanding to the task-specific design. The machine that runs the programmer's program does not understand the handcrafted symbolic model, since the reasons for the symbolic design are in the mind of the human designer, but not understood by the machine. The machine is not able to learn new concepts beyond a finite number of combinations of hand-selected symbolic concepts. As we discussed earlier, there is an exponential number of sensory states that need to be distinguished in a real world environment, but only a moderate number of symbols that a human can handcraft and program. This seems to be the major reason that a symbolic model is brittle in the real world.

B. Emergent

Cresceptron by Weng et al. 1992 [93], [94] appears to be the first published visual learning self-organizing network for recognizing and segment general objects from natural backgrounds. The open problems that Cresceptron left to us include optimal use of a limited number of neurons, optimal use of learning experience, how to learn multiple concepts concurrently without handcrafting any concept (e.g., not just design a network for the type concept but use convolution to forget location), top-down attention, goal emergence, and goal directed perception. The theoretical work here addresses this array of problems using a cortex-inspired and high integrated architecture and representation, with a discussion of a series of experimental results.

There have been several versions of networks that use error back-propagation [103], [44], [22], [104]. We do not think that the brain uses error back-propagation since there seems no such biological evidence and further motor errors are not available during autonomous practice and self-learning through exploration. All error back-propagation networks, including the versions that freeze part of the network [22], do not have an effective method to dynamically distinguish working memory from long-term memory while dynamically maintaining a bounded amount of memory resource as LCA does without "running out of memory".

Olshaushen & Field [59] published an interesting computational work. They started from an objective function (cost function) to minimize the image reconstruction error from a set of (synaptic) weight vectors while encouraging fewer weight vectors to contribute to the reconstruction of each image. They used the learning rule derived from the objective function to determine the weight vectors using many natural images. Interestingly, their resulting weight vectors look spatially local and oriented [59, Fig. 4a], like many local edge detectors.

Our model supports the sparse coding idea of Olshaushen & Field [59] — each cortical area has few neurons to fire. But our model is different in a number of important aspects. Brain's sparse internal coding is not just a style of over-complete representation as some researchers suggested, but more importantly, it affords the need for competition so that only the best matched neurons are allowed to fire and update, to avoid erasing long term memory. Instead of starting from a hypothetical objective function, our model starts from the cortical connection patterns experimentally observed in many physiological studies, as surveyed and summarized by Felleman & Van Essen [25]. The learning equation in our

model is consistent with the experimentally observed Spike Timing Dependent Plasticity (STDP) [7], [15], different from the learning equation of Olshaushen & Field [59] which requires hypothetically reconstructed images.

It seems reasonable to expect that *evolutionary pressure on the brain is about the behaviors from the brain, not directly for reconstructing images.* Our model does not consider sparse coding as an objective but rather a result from the existence of lateral inhibitory connections (simulated by top-k competition) widely found in the cortex. The competition enables most neurons to be stable (not firing and updating) at any time to keep their weights as long-term memory, as explained in Sec. II-K. The most fundamental point of our model, different from Olshaushen & Field's work is that each brain area uses not only bottom-up input, but top-down input conjunctively for behavior generation (not really for image reconstruction).

Tenenbaum et al. [84] published a technique, called Isomap, inspired by the need for the brain to reduce the dimension d (e.g., the number of optical nerves from the retina) of sensory inputs. Their idea is to find a lower dimensional space in which the geodesic distance between sensory data points in the original space is maintained in the lower dimensional space. They assume that the sensory data points lie in a single continuous lower dimensional nonlinear manifold (e.g., on the surface of a Swiss roll). In contrast, each motor neuron in the WWN here automatically "finds", using simple Hebbian learning, a complex manifold of neurons in the internal Y space which is not necessarily connected as Isomap requires. Nor does WWN spend expensive computation to compute the eigenvalues, eigenvectors, and geodesic distance that Isomap requires.

The Locally Linear Embedding (LLE) by Roweis and Saul [70] is similar to Isomap in terms of motivation, but uses a discrete derivation which leads to a computational procedure similar to the Isomap in principle, also computing the eigenvalues and eigenvectors of a monolithic data matrix, which does not use motor information. In contrast, the computation of WWN is incremental, in-place, and discriminative by using motor information, in addition to the advantages in the above comparison with Isomap.

The reduction of sensory dimension is necessary. Isomap and LLE are two well known methods that maintain within-manifold distance. Our model also reduces the sensory dimension, but it differs from Isomap and LLE in a number of ways. The goal of Isomap and LLE is to represent the sensory space. Our model finds an optimal bridge representation for both bottom and top spaces, i.e., $X \times Z$, for the goal to generate *desired behaviors*, not primarily for representing the *sensory space*. Isomap and LLE use eigenvectors to represent features, which do not give sparse coding.

The Deep Belief Networks (DBN) of LeCun et al. 1998 [44], Mumford 2003 [45], Hinton et al. 2006 [34], use a cascade of areas. They have not been shown to detect objects from complex backgrounds. As discussed above, WWN does not use a cascade of areas.

Isomap, LLE, the Olshaushen-Field features, and DBN are all unsupervised, with a primary goal of learning internal presentation that can reconstruct input image well. The primary goal of our brain-inspired model is to generate behaviors. The new WWN uses supervised leaning and self-learning (learn through self-practice and exploration as it generates its own goals).

In terms of image input, the representation schemes of the Olshaushen-Field features, Isomap, LLE, and DBN all use *global* receptive field (entire retina) for every feature neuron. A few other recurrent networks [68], [76] and the ancestor of WWN, the recurrent Multilayer In-place Learning Networks (MILN) [98], [99] also used global receptive field. Our model here is different, as WWN uses a *local* receptive field for each Y neuron, so that each individual object in a complex background is reported individually by the best matched neuron among those who have roughly the correct receptive field. The automatic top-down wiring is critical for a goal in Z to perform goal-directed perception inside the network in the presence of multiple objects.

VI. Conclusions and Discussion

The theory, model, and algorithm presented in this paper show that if the brain B is properly trained, a subpart in the sensory end S (e.g., an object view regardless of the background) and/or a subpart in the motor end (e.g., a particular type goal or location goal) is sufficient to regulate the competition in brain B which allows the best matched Y neuron for S and M to fire. In the next network update, this firing Y neuron, as the pre-synaptic firing, predicts the firing for all the post-synaptic neurons in Z as the corresponding action in M (also predicts the firing in S if S computes like M).

In cognitive science and AI, symbolic (abstract), connectionist (concrete), and behaviorist (acting) directions have their strengths, although they collectively are not sufficient to close the gap. It has been established [90] that given any FA, a DN can learn such an FA incrementally through interactions by representing the state as its motor output. An FA is symbolic, handcrafted and static once designed without any internal representation. In contrast, the DN is incrementally taught and can autonomously practice and thus self-learn by autonomously self-organizing its internal representations. With the DN theory here and FA as a special case, we propose that the age where connectionist models categorically cannot perform goal-directed reasoning (including perception, recognition and search) is over. Much exciting future lies ahead. For example, it will be interesting to see how the model scales up like the cerebral cortex.

Biologically, the work here predicts: Each neuron is not pure linguistically in any human language due to its six diverse sources of input. This is in contrast with all symbolic models and symbolic-emergent hybrid models, which impose rigid walls in the system between different linguistically expressed meanings (symbols). Likewise, the function of any brain area cannot be stated precisely using any concept of the extra-body environment. This is in contrast with a common practice that states that a brain area does X (e.g., detecting oriented edges — an extra-body concept). Any human communicable concept may be supervised at a corresponding motor area (e.g., verbal

teaching), since it is a port where external teachers can sense, influence, and calibrate.

Genetically, cellular mechanisms (e.g., the Hebbian mechanism and the cellular scheduling of plasticity) seem sufficient, in principle, to wire up through experience a sophisticated "brain" that demonstrates abstract behaviors from concrete receptors and muxels. Not being a holistically-aware central controller, the DP of DN is sufficient to regulate the hextuple network representations that enable goal-directed attention, goal-perception and goal-recognition from pixels.

REFERENCES

[1] M. Albanese, R. Chellappa, N. Cuntoor, V. Moscato, A. Picariello, A. Picariello, and O. Udrea. PADS: A probabilistic activity detection framework for video data. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(12):2246–2261, 2010.

[2] N. Almassy, G. M. Edelman, and O. Sporns. Behavioral constraints in the development of neural properties: A cortical model embedded in a real-world device. *Cerebral Cortex*, 8(4):346–361, 1998.

[3] C. H. Anderson and D. C. Van Essen. Shifter circuits: A computational strategy for dynamic aspects of visual processing. *Proc. Natl. Acad. Sci. USA*, 84:6297–6301, Sept. 1987.

[4] J. R. Anderson. *Rules of the Mind*. Lawrence Erlbaum, Mahwah, New Jersey, 1993.

[5] D. Ansari. Effects of development and enculturation on number representation in the brain. *Nature Reviews Neuroscience*, 9:278–291, 2008.

[6] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida. Cognitive developmental robotics: A survey. *IEEE Trans. Autonomous Mental Development*, 1(1):12–34, 2009.

[7] G. Bi and M. Poo. Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience*, 24:139–166, 2001.

[8] C. Blakemore and G. F. Cooper. Development of the brain depends on the visual environment. *Nature*, 228:477–478, Oct. 1970.

[9] E. M. Callaway. Local circuits in primary visual cortex of the macaque monkey. *Annual Review of Neuroscience*, 21:47–74, 1998.

[10] S. Carey. Cognitive development. In D. N. Osherson and E. E. Smith, editors, *Thinking*, pages 147–172. MIT Press, Cambridge, Massachusetts, 1990.

[11] S. Carey. Précis of the origin of concepts. *Behavioral and Brain Science*, 34:113–167, 2011.

[12] M. Cole and S. R. Cole. *The Development of Children*. Freeman, New York, 3rd edition, 1996.

[13] M. Corbetta and G. L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neural Science*, 3:201–215, 2002.

[14] M. C. Crair, D. C. Gillespie, and M. P. Stryker. The role of visual experience in the development of columns in cat visual cortex. *Science*, 279:566–570, 1998.

[15] Y. Dan and M. Poo. Spike timing-dependent plasticity: From synapses to perception. *Physiological Review*, 86:1033–1048, 2006.

[16] G. Deco and E. T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 40:2845–2859, 2004.

[17] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222, 1995.

[18] M. Domjan. *The Principles of Learning and Behavior*. Brooks/Cole, Belmont, California, fourth edition, 1998.

[19] R. J. Douglas and K. A. C. Martin. Neural circuits of the neocortex. *Annu. Rev. Neurosci.*, 27:419–451, 2004.

[20] M. P. Dubuisson, J. S. Lakshmanan, and A. K. Jain. Vehicle segmentation and classification using deformable templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(3):293–308, 1996.

[21] J. L. Elman, E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett. *Rethinking Innateness: A connectionist perspective on development*. MIT Press, Cambridge, Massachusetts, 1997.

[22] S. E. Fahlman and C. Lebiere. The cascade-correlation learning architecture. Technical Report CMU-CS-90-100, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, Feb. 1990.

[23] A. Fazl, S. Grossberg, and E. Mingolla. View-invariant object category learning, recognition, and search: How spatial and object attention are coordinated using surface-based attentional shrouds. *Cognitive Psychology*, 58:1–48, 2009.

[24] L. Fei-Fei. One-shot learning of object categories. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[25] D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.

[26] M. B. Feller, D. P. Wellis, D. Stellwagen, F. S. Werblin, and C. J. Shatz. Requirement for cholinergic synaptic transmission in the propagation of spontaneous retinal waves. *Science*, 272(5265):1182–1187, 1996.

[27] D. Fitzpatrick. Seeing beyond the receptive field in primary visual cortex. *Current Opinion in Neurobiology*, 10(4):438–443, 2000.

[28] J. H. Flavell. Cognitive development: Past, present, and future. In K. Lee, editor, *Child Cognitive Development*, pages 7–29. Blackwell, Maden, Massachusetts, 2000.

[29] J. H. Flavell, P. H. Miller, and S. A. Miller. *Cognitive Development*. Prentice Hall, New Jersey, 3rd edition, 1993.

[30] M. D. Fox, M. Corbetta, A. Z. Snyder, J. L. Vincent, and M. E. Raichle. Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proc. National Academy of Sciences U S A*, 103(26):10046–10051, 2006.

[31] M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15:20–25, 1992.

[32] J. T. Green, R. B. Ivry, and D. S. Woodruff-Pak. Timing in eyeblink classical conditioning and timed-interval tapping. *Psychological Science*, 10(1):19–25, 1999.

[33] S. Grossberg and R. Raizada. Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Research*, 40:1413–1432, 2000.

[34] G. E. Hinton, S. Osindero, and Y-. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[35] F. H. Hsu. IBM's deep blue chess grandmaster chips. *IEEE Micro*, 19(2):70–81, 1999.

[36] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, 2001.

[37] L. Itti, G. Rees, and J. K. Tsotsos, editors. *Neurobiology of Attention*. Elsevier Academic, Burlington, MA, 2005.

[38] Z. Ji and J. Weng. WWN-2: A biologically inspired neural network for concurrent visual attention and recognition. In *Proc. IEEE Int'l Joint Conference on Neural Networks*, pages +1–8, Barcelona, Spain, July 18-23 2010.

[39] Z. Ji, J. Weng, and D. Prokhorov. Where-what network 1: "Where" and "What" assist each other through top-down connections. In *Proc. IEEE Int'l Conference on Development and Learning*, pages 61–66, Monterey, CA, Aug. 9-12 2008.

[40] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, editors. *Principles of Neural Science*. McGraw-Hill, New York, 4th edition, 2000.

[41] E. I. Knudsen. Fundamental components of attention. *Annual Reviews Neuroscience*, 30:57–78, 2007.

[42] C. Koch. Being john malkovich: Personal control of individual brain cells. *Scientific American*, pages 18–19, March/April 2011.

[43] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[44] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of IEEE*, 86(11):2278–2324, 1998.

[45] T. S. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *J. Opt. Soc. Am. A*, 20(7):1434–1448, 2003.

[46] T. S. Lee, C. F. Yang, R. D. Romero, and D. Mumford. Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency. *Nature Neuroscience*, 5(6):589–597, 2002.

[47] W. R. Lippe. Rhythmic spontaneous activity in the developing avian auditary system. *Journal of Neuroscience*, 14(3):1486–1495, 1994.

[48] N. K. Logothetis and D. L. Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19:577–621, 1996.

[49] M. Luciw and J. Weng. Laterally connected lobe component analysis: Precision and topography. In *Proc. IEEE 8th Int'l Conference on Development and Learning*, pages +1–8, Shanghai, China, June 4-7 2009.

[50] M. Luciw and J. Weng. Top-down connections in self-organizing Hebbian networks: Topographic class grouping. *IEEE Trans. Autonomous Mental Development*, 2(3):248–261, 2010.

[51] M. Luciw and J. Weng. Where What Network 3: Developmental top-down attention with multiple meaningful foregrounds. In *Proc. IEEE Int'l Joint Conference on Neural Networks*, pages 4233–4240, Barcelona, Spain, July 18-23 2010.

[52] M. Luciw, J. Weng, and S. Zeng. Motor initiated expectation through top-down connections as abstract context in a physical world. In *IEEE Int'l Conference on Development and Learning*, pages +1–6, Monterey, CA, Aug. 9-12 2008.

[53] R. Miikkulainen, J. A. Bednar, Y. Choe, and J. Sirosh. *Computational Maps in the Visual Cortex*. Springer, Berlin, 2005.

[54] M. Minsky. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI Magazine*, 12(2):34–51, 1991.

[55] M. Mishkin, L. G. Unterleider, and K. A. Macko. Object vision and space vision: Two cortical pathways. *Trends in Neuroscicence*, 6:414–417, 1983.

[56] K. Miyan and J. Weng. WWN-Text: Cortex-like language acquisition with What and Where. In *Proc. IEEE 9th Int'l Conference on Development and Learning*, pages 280–285, Ann Arbor, August 18-21 2010.

[57] M. J. O'Donovan. The origin of spontaneous activity in developing networks of the vertebrate nervous systems. *Current Opinion in Neurobiology*, 9:94–104, 1999.

[58] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13(11):4700–4719, 1993.

[59] B. A. Olshaushen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, June 13 1996.

[60] L. M. Parsons. Imagined spatial transformations of ones hands and feet. *Cognitive Psychology*, 19:178–241, 1987.

[61] J. Piaget. *The Construction of Reality in the Child*. Basic Books, New York, 1954.

[62] W. K. Purves, D. Sadava, G. H. Orians, and H. C. Heller. *Life: The Science of Biology*. Sinauer, Sunderland, MA, 7 edition, 2004.

[63] S. Quartz and T. J. Sejnowski. The neural basis of cognitive development: A constructivist manifesto. *Behavioral and Brain Sciences*, 20(4):537–596, 1997.

[64] R. Q. Quiroga, A. Kraskov, 2 C. Koch, and I. Fried. Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology*, 19(1):13081313, 2009.

[65] R. Rao and D. H. Ballard. An active vision architecture based on iconic representation. *Artificial Intelligence*, 78:461–505, 1995.

[66] A. S. Reber, S. M. Kassin, S. Lewis, and G. Cantor. On the relationship between implicit and explicit modes in the learning of a complex rule structure. *Journal of Experimental Psychology: Human Learning and Memory*, 6(5):492–502, 1980.

[67] M. Riesenhuber and T. Poggio. Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, 12(2):162168, 2002.

[68] P. R. Roelfsema and A. van Ooyen. Attention-gated reinforcement learning of internal representations for classification. *Neural Computation*, 17:2176–2214, 2005.

[69] P. S. Rosenbloom, J. E. Laird, and A. Newell, editors. *The Soar Papers*. MIT Press, Cambridge, Massachusetts, 1993.

[70] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(2319):2323–2326, 2000.

[71] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Upper Saddle River, New Jersey, 2nd edition, 2003.

[72] T. J. Sejnowski. What are the projective fields of cortical neurons? In L. J. van Hemmen and T. J. Sejnowski, editors, *Twenty Three Problems in Systems Neuroscience*, page 394405. Oxford University Press, Oxford, 2006.

[73] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007.

[74] J. Sharma, A. Angelucci, and M. Sur. Induction of visual orientation modules in auditory cortex. *Nature*, 404:841–847, 2000.

[75] R. N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701–703, 1971.

[76] Y. F. Sit and R. Miikkulainen. Self-organization of hierarchical visual maps with feedback connections. *Neurocomputing*, 69:1309–1312, 2006.

[77] M. Solgi and J. Weng. Developmental stereo: Emergence of disparity preference in models of visual cortex. *IEEE Trans. Autonomous Mental Development*, 1(4):238–252, 2009.

[78] X. Song, W. Zhang, and J. Weng. Where-what network 5: Dealing with scales for objects in complex backgrounds. In *Proc. Int'l Joint Conference on Neural Networks*, pages +1–8, San Jose, CA, July 31 - August 5 2011.

[79] A. B. Steinmetz and C. R. Edward. Comparison of auditory and visual conditioning stimuli in delay eyeblink conditioning in healthy young adults. *Learning and Behavior*, 37:349–356, 2009.

[80] R. Sun, P. Slusarz, and C. Terry. The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, 112(1):59192, 2005.

[81] R. Sun and X. Zhang. Accounting for a variety of reasoning data within a cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, 18:169–191, 2006.

[82] M. Sur, A. Angelucci, and J. Sharm. Rewiring cortex: The role of patterned activity in development and plasticity of neocortical circuits. *Journal of Neurobiology*, 41:33–43, 1999.

[83] M. Sur and J. L. R. Rubenstein. Patterning and plasticity of the cerebral cortex. *Science*, 310:805–810, 2005.

[84] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[85] A. M. Treisman. A feature-integration theory of attention. *Cognitive Science*, 12(1):97–136, 1980.

[86] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78:507–545, 1995.

[87] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *Int'l J. of Computer Vision*, 63(2):113–140, 2005.

[88] Y. Wang, X. Wu, and J. Weng. Synapse maintenance in the where-what network. In *Proc. Int'l Joint Conference on Neural Networks*, pages +1–8, San Jose, CA, July 31 - August 5 2011.

[89] J. Weng. Task muddiness, intelligence metrics, and the necessity of autonomous mental development. *Minds and Machines*, 19(1):93–115, 2009.

[90] J. Weng. A 5-chunk developmental brain-mind network model for multiple events in complex backgrounds. In *Proc. Int'l Joint Conf. Neural Networks*, pages 1–8, Barcelona, Spain, July 18-23 2010.

[91] J. Weng. Three theorems: Brain-like networks logically reason and optimally generalize. In *Proc. Int'l Joint Conference on Neural Networks*, pages +1–8, San Jose, CA, July 31 - August 5 2011.

[92] J. Weng. Symbolic models and emergent models: A review. *IEEE Trans. Autonomous Mental Development*, 3:+1–26, 2012. Accepted and to appear.

[93] J. Weng, N. Ahuja, and T. S. Huang. Cresceptron: a self-organizing neural network which grows adaptively. In *Proc. Int'l Joint Conference on Neural Networks*, volume 1, pages 576–581, Baltimore, Maryland, June 1992.

[94] J. Weng, N. Ahuja, and T. S. Huang. Learning recognition and segmentation using the Cresceptron. *International Journal of Computer Vision*, 25(2):109–143, Nov. 1997.

[95] J. Weng and S. Chen. Visual learning with navigation as an example. *IEEE Intelligent Systems*, 15:63–71, Sept./Oct. 2000.

[96] J. Weng and W. Hwang. From neural networks to the brain: Autonomous mental development. *IEEE Computational Intelligence Magazine*, 1(3):15–31, 2006.

[97] J. Weng and M. Luciw. Dually optimal neuronal layers: Lobe component analysis. *IEEE Trans. Autonomous Mental Development*, 1(1):68–85, 2009.

[98] J. Weng and M. D. Luciw. Optimal in-place self-organization for cortical development: Limited cells, sparse coding and cortical topography. In *Proc. 5th Int'l Conference on Development and Learning (ICDL'06)*, pages +1–7, Bloomington, IN, May 31 - June 3 2006.

[99] J. Weng, T. Luwang, H. Lu, and X. Xue. Multilayer in-place learning networks for modeling functional layers in the laminar cortex. *Neural Networks*, 21:150–159, 2008.

[100] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen. Autonomous mental development by robots and animals. *Science*, 291(5504):599–600, 2001.

[101] J. Weng and N. Zhang. Optimal in-place learning and the lobe component analysis. In *Proc. IEEE World Congress on Computational Intelligence*, pages +1–8, Vancouver, BC, Canada, July 16-21 2006.

[102] J. Weng, Q. Zhang, M. Chi, and X. Xue. Complex text processing by the temporal context machines. In *Proc. IEEE 8th Int'l Conference on Development and Learning*, pages +1–8, Shanghai, China, June 4-7 2009.

[103] P. J. Werbos. *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*. Wiley, Chichester, 1994.

[104] G. Westermann, S. Sirois, T. R. Shultz, and D. Mareschal. Modeling developmental cognitive neuroscience. *Trends in Cognitive Sciences*, 10(5):227–232, 2006.

[105] A. K. Wiser and E. M. Callaway. Contributions of individual layer 6 pyramidal neurons to local circuitry in macaque primary visual cortex. *Journal of neuroscience*, 16:2724–2739, 1996.

[106] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *Proc. Computer Vision*

and Pattern Recognition, pages +1–8, San Francisco, CA, June 15-17 2010.

[107] Y. C. Yu, R. S. Bultje, X. Wang, and S. H. Shi. Specific synapses develop preferentially among sister excitatory neurons in the neocortex. *Nature*, 458(7237):501–504, 2009.

Juyang Weng (S85-M88-SM05-F09) received the BS degree in computer science from Fudan University, Shanghai, China, in 1982, and M. Sc. and PhD degrees in computer science from the University of Illinois at Urbana-Champaign,in 1985 and 1989, respectively.

He is currently a professor of Computer Science and Engineering at Michigan State University, East Lansing. He is also a faculty member of the Cognitive Science Program and the Neuroscience Program at Michigan State University. Since the work of Cresceptron (ICCV 1993), he expanded his research interests in biologically inspired systems, especially the autonomous development of a variety of mental capabilities by robots and animals, including perception, cognition, behaviors, motivation, and abstract reasoning skills. He has published over 250 research articles on related subjects, including task muddiness, intelligence metrics, mental architectures, vision, audition, touch, attention, recognition, autonomous navigation, natural language understanding, and other emergent behaviors.

Dr. Weng is an Editor-in-Chief of International Journal of Humanoid Robotics and an associate editor of the IEEE Transactions on Autonomous Mental Development. He was a Program Chairman of the NSF/DARPA funded Workshop on Development and Learning 2000 (1st ICDL), a Program Chairman of the 2nd ICDL (2002), the chairman of the Autonomous Mental Development Technical Committee of the IEEE Computational Intelligence Society (2004-2005), the Chairman of the Governing Board of the International Conferences on Development and Learning (ICDLs) (2005-2007), a General Chairman of 7th ICDL (2008), the General Chairman of 8th ICDL (2009), an associate editor of IEEE Transactions on Pattern Recognition and Machine Intelligence, and an associate editor of IEEE Transactions on Image Processing.

Matthew Luciw received the M.S. and Ph.D. degrees in computer science from Michigan State University (MSU), East Lansing, in 2006 and 2010, respectively.

He was previously a member of the Embodied Intelligence Laboratory at MSU. He is currently working as a researcher at the Dalle Molle Institute for Artificial Intelligence (IDSIA), Manno-Lugano, Switzerland. His research involves the study of biologically-inspired algorithms to enable autonomous learning agents. He is a member of the IEEE Computational Intelligence Society.