

# Brain-Inspired Concept Networks: Learning Concepts from Cluttered Scenes

Juyang Weng, *Michigan State University*

Matthew D. Luciw, *Boston University*

**A**s defined in Webster's dictionary, a concept is an abstract or generic idea generalized from particular instances. For example, object location (for example, in an image) is a concept that can be learned from many particular instances of various objects so that it becomes type invariant. Likewise,

the concept of object type (for example, in a cluttered scene) is a concept that can be learned from many particular instances of various locations of the same type so that it becomes location invariant. In general, a concept and its value is represented as a sequence of response patterns in the motor area, typically in a natural language, to represent concepts such as location, type, scale, goal, subgoal, intent, purpose, price, and so on.

Although the discussion here is hopefully applicable to various sensory and effector modalities, we focus on vision as a sensory modality and only two categories of concepts: location and type. Other concepts are similar (for example, scale concept as in work by Xiaoying Song and her colleagues<sup>1</sup>). The reader can notice that the mechanisms discussed later aren't specific to either the sensory modality or the two chosen motor modality categories.

There are two types of visual attention: bottom-up and top-down.<sup>2-5</sup> The bottom-up process is largely intent-free (or concept-free), which is evident in a free-viewing situation when objects compete to catch our attention and a winner pops up.<sup>6</sup> However, the top-down process is concept-directed. The network circuitry for the top-down process has been elusive. Robert Desimone and John Duncan wrote: "So far we have not dealt specifically with the representation of objects in the cortex. Although this is a key issue for understanding attention, little is actually known about the neural representations of objects."<sup>2</sup> The lack of computational modeling of a network's internal object representation has continued (for example, see a review by Eric I. Knudsen<sup>7</sup>).

Charles H. Anderson,<sup>8</sup> Bruno A. Olshausen,<sup>9</sup> and John K. Tsotsos and their colleagues<sup>10</sup> proposed shifter circuits for location-based, top-down attention (that is, not including the

*Inspired by the anatomical connection patterns in the cerebral cortex, the authors introduce concept networks. Such a network acquires concepts as actions through autonomous, incremental, and optimal self-wiring and adaptation.*

type-based, top-down attention in this work). Given a location and scale (that is, given the location and scale values), their circuits shift and scale the selected subpart in the retinal image into a size-normalized and background-free master map. From this master map, a subsequent classifier produces a label of object type. The values of object location and scale are from a separate mechanism, not from the same network. In contrast, our concept network doesn't have such a master map because such a map is handcrafted by the human programmer (as an internal central controller) for the given location concept only. We don't think that such a central controller exists for any biological brain.

There are many computational symbolic models<sup>11</sup> in the computer vision community, but they use handcrafted 3D object models or 2D appearance models, which don't allow the system to create unmodeled concepts. See Juyang Weng's work for the fundamental limitations of symbolic models.<sup>11</sup>

Our concept network has embodiments called where-what networks.<sup>12</sup> A where-what network has learned at least two concepts, location (where) and type (what). It provides a highly integrated computational model that unifies all three types of attention—bottom-up; location-based, top-down; and type-based, top-down. The concepts and the circuits for executing the ideas emerge from, and are embedded in, the same network. A where-what network also gives a schematic and unified solution to the emergent internal representation for which Robert Desimone and his colleague<sup>2</sup> and Eric I. Knudsen<sup>7</sup> called. This solution indicates that a master map isn't necessary for all three types of attention. Furthermore, a concept network provides a unified solution to the three open problems in the right three columns of Table 1, which compares symbolic

Table 1. Comparison of models.

| Problems addressed      | Goal-directed reasoning | Emergent representation | Unmodeled concepts | Concept emergence | Concept-directed perception |
|-------------------------|-------------------------|-------------------------|--------------------|-------------------|-----------------------------|
| Symbolic models         | Yes                     | No                      | No                 | No                | No                          |
| Prior emergent networks | No                      | Yes                     | No                 | No                | No                          |
| Concept networks        | Yes                     | Yes                     | Yes                | Yes               | Yes                         |

models, prior emergent networks (neural nets), and concept networks.

With limited learning experience, our experimental results reported here only deal with early concepts in life (that is, location and type).

### Concept Networks

Now, let's take a closer look at the concept networks.

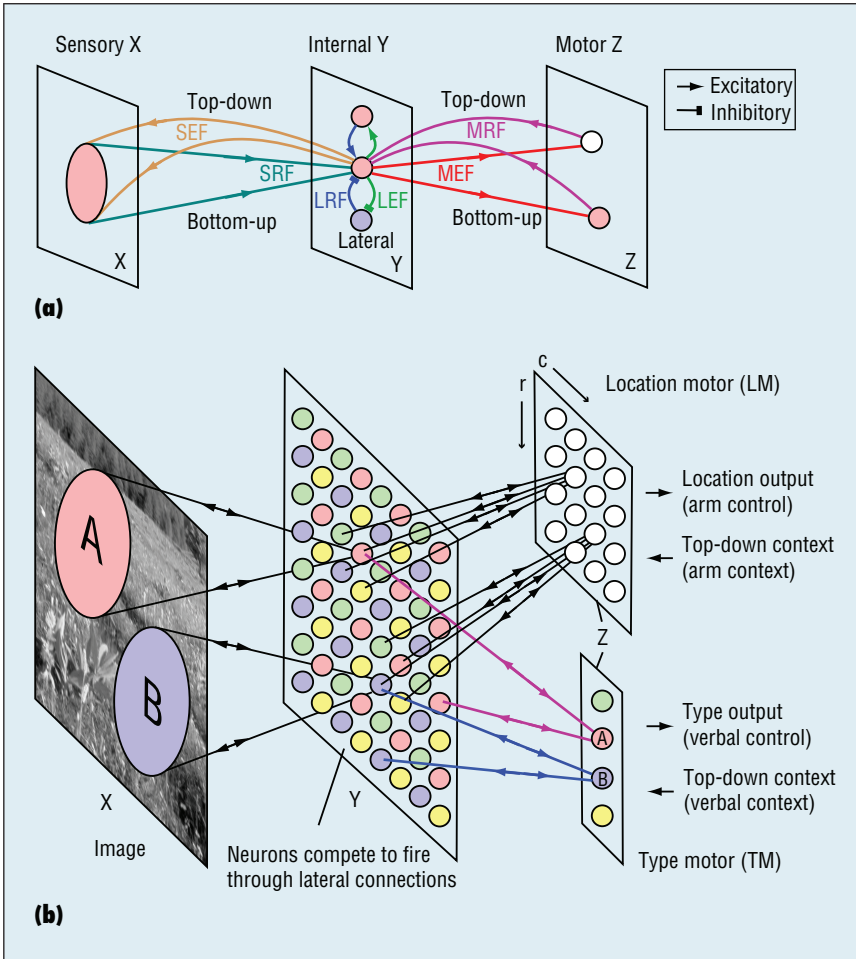
#### Theory

More than 50 neuroanatomical studies reviewed by Daniel J Felleman and David Van Essen indicated that each cortical area receives not only bottom-up (ascending) input from earlier (sensory) processing areas but also top-down (descending) input from later (motoric) processing areas.<sup>13</sup> Furthermore, each area's connections to earlier and later areas are both two-way. Among neurons in the same area, there are also intra-area (lateral) connections. That is, each neuron in general has three sources of inputs: bottom-up, lateral, and top-down, as Figure 1a illustrates. Such ubiquitous recurrence has posed great challenges to circuit understanding and analysis. Some prior two-way networks<sup>14</sup> drop all intra-area (lateral) connections to simplify the processing. For more discussion about support from brain studies, see Juyang Weng's work.<sup>12</sup>

Each concept network has two external areas  $X$  and  $Z$  and one (large) internal area  $Y$ , as Figure 1b shows. Each area  $A$  in  $\{X, Y, Z\}$  as a bridge predicts the signals in its two islands (sensory area and motor area) using its limited resource, as Figure 1a

shows. If  $Y$  is the entire nervous system, then  $X$  consists of all sensory receptors (for example, retina), and  $Z$  consists of all the motor neurons that drive muscles and glands. All levels in  $Y$  are emergent. The concept network here computationally models how each brain area (for example, a Brodmann area) predicts signals in two island areas to which it connects.  $X$  and  $Z$  serve as both input and output: environment teaching patterns go into  $Z$  (that is, the "Emergent representation" column of Table 1) and, later,  $Y$  and  $X$  are predicted into  $Z$ , and unmodeled concepts (goals, actions, and so on) emerge from  $Z$  (that is, the "Unmodeled concepts" column of Table 1). Camera-supervising patterns go into  $X$  and, later,  $Z$  and  $Y$  are predicted into  $X$ , and the network executes attention/prediction (that is, the "Concept emergence" column of Table 1).

As Figure 1b models, within the internal brain area  $Y$ , each neuron connects with highly correlated neurons using excitatory connections (for example, via N-methyl-D-aspartate receptors) but connect with highly anticorrelated neurons using inhibitory connections (for example, via gamma-aminobutyric acid receptors). This forces neurons in the same area to detect different features in the sensory receptive field and motor receptive field. Suppose that location motor (LM) and type motor (TM) are always taught with the correct object location and object type, respectively. Suppose also that an early developed pulvinar area only allows



**Figure 1.** The architecture of concept networks. (a) Six fields (hextuple) of each neuron: sensory receptive field (SRF), motor receptive field (MRF), lateral receptive field (LRF), sensory effective field (SEF), motor effective field (MEF), and lateral effective field (LEF). (b) A simple concept network with four areas: image X, occipital lobe Y, and location motor and type motor in the frontal lobe Z. Each wire connects if the presynaptic and postsynaptic neurons have cofired. A two-way arrow means two one-way connections. All the lateral connections within the same area are omitted for clarity: at this sparse neuronal density only lateral inhibitory connections survive, and they connect with all neuronal pairs in Y.

those object-looking Y neurons to fire and learn, but not background looking neurons to avoid wasting many Y neurons to learn background patterns. The object-looking Y neurons are those neurons whose sensory receptive fields are at the object location signaled by LM. Then, these developmental mechanisms result in the automatically generated connections in Figure 1 via Hebbian learning: Every Y neuron is location and type specific for a taught object. Each LM neuron is location specific for a taught

location but type invariant for all taught types. Each TM neuron is type specific for a taught type but location invariant for all taught locations. Each Z motor neuron bottom-up collects all applicable cases (neurons) from Y, each Y neuron being a case for the Z neuron to fire. The Z neuron also top-down boosts all applicable cases (neurons) down to Y (for example, each TM neuron type boosts all applicable Y neurons of the attended type so it can find the object in a cluttered scene). The location concept in

LM and the location concept in TM are all taught by the environment, and aren't a part of programming. The internal wiring is determined by the statistical nature of the concept taught in LM and TM, respectively, as well as the contents in the image X.

**Live**

During prenatal learning, the  $c$  neurons of Y need to initialize their synaptic vectors  $V = (v_1, v_2, \dots, v_c)$ , and the firing ages  $A = (n_1, n_2, \dots, n_c)$ . Each synaptic vector  $v_i$  is initialized using the input pair  $p_i = (x_i, z_i)$ ,  $i = 1, 2, \dots, c$ , at the first  $c$  time instances. Each firing age  $a_i$  is initialized to be zero,  $i = 1, 2, \dots, c$ .

After birth, at each time instant, each area A in {X, Y, Z} computes its response  $r'$  from its bottom-up input  $b$  and top-down input  $t$  with  $p = (b, t)$  based on its adaptive part  $N = (V, A)$  and its current response  $r$ , background suppressing vector  $r_a$  (used only by Y) from LM, and updates  $N$  to  $N'$ :

$$(r', N') = f(b, r, t, r_a, N), \tag{1}$$

where  $f$  is the unified area function described around Equations 2 through 4. The vector  $r_a$  has the same dimension as  $r$ , suppresses all the Y neurons to zeros except the  $3 \times 3 = 9$  ones centered at the correct object location. The vector  $r_a$  isn't available when LM isn't supervised (for example, during testing sessions).

Note that sensory area X doesn't have a bottom-up area, and the motor area Z doesn't have a top-down area. However, they predict the next firing according to the supervision input from the environment and the firing pattern in Y.

**Area Function**

We hypothesize that each brain area performs prediction for all these

connected areas through its two-way connections. It seems then desirable that different areas share the same form of area functions and the same form of learning mechanisms. Of course, every area performs a different type of prediction because each area function is different as a result of its own learning. The concept network is highly recurrent. It's important that only a few neurons in each area fire and update so that those that don't update serve as long-term memory for the current context. The interneuron competition is based on the goodness of the match between a neuron's weight vector  $\mathbf{v} = (\mathbf{v}_b, \mathbf{v}_t)$ . Let  $\hat{\mathbf{x}}$  denote  $\mathbf{x}/\|\mathbf{x}\|$ , the unit version of vector  $\mathbf{x}$ . The goodness of a match is measured by the preaction potential:

$$r(\mathbf{v}_b, \mathbf{b}, \mathbf{v}_t, \mathbf{t}) = \hat{\mathbf{v}} \cdot \hat{\mathbf{p}}, \quad (2)$$

where  $\mathbf{v} = (\hat{\mathbf{v}}_b, \hat{\mathbf{v}}_t)$ , and  $\mathbf{p} = (\hat{\mathbf{b}}, \hat{\mathbf{t}})$ . The inner product measures the degree of match between the two directions  $\hat{\mathbf{v}}$  and  $\hat{\mathbf{p}}$ , because  $r(\mathbf{v}_b, \mathbf{b}, \mathbf{v}_t, \mathbf{t}) = \cos(\theta)$ , where  $\theta$  is the angle between two unit vectors  $\hat{\mathbf{v}}$  and  $\hat{\mathbf{p}}$ . This enables a match between two vectors of different magnitudes (for example, a weight vector from an object viewed indoor to match the same object when it's viewed outdoor). The preresponse value ranges are  $[-1, 1]$ .

It's known in electrical engineering that positive feedbacks may cause uncontrollable oscillations and system instability. Our computational theory for a cortical area, the lobe component analysis,<sup>15</sup> uses a top- $k$  firing mechanism—a highly nonlinear mechanism—to explain that lateral inhibitions enable neurons in each area  $Y$  to sort out top- $k$  winners within each time step  $t_n$ ,  $n = 1, 2, 3, \dots$ . Let the weight vector of neuron  $i$  be  $\mathbf{v}_i = (\mathbf{v}_{bi}, \mathbf{v}_{ti})$ ,  $j = 1, 2, \dots, c$ , where  $\mathbf{v}_{bi}$  and  $\mathbf{v}_{ti}$  are the weight vectors of the bottom-up input  $\mathbf{b}$  and top-down

input  $\mathbf{t}$  of neuron  $i$ , respectively. See our related work for the effects of  $k$ .<sup>15</sup> For simplicity, considering  $k = 1$ , the single winner neuron  $j$  is identified by

$$j = \arg \max_{1 \leq i \leq c} r(\mathbf{v}_{bi}, \mathbf{b}, \mathbf{v}_{ti}, \mathbf{t}). \quad (3)$$

Suppose  $c$  is sufficiently large and the set of  $c$  synaptic vectors distributes well. Then, with sufficient training (that is, data epochs), the winner (nearest-neighbor) neuron  $j$  has both of its parts match well

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

not counting the lengths of these vectors because of the previously discussed normalization in computing  $r(\mathbf{v}_b, \mathbf{b}, \mathbf{v}_t, \mathbf{t})$ . This is like logic-AND: both parts must match well for the neuron  $j$  to win.

We would like to have the response value  $r_j$  to approximate the probability for  $(\mathbf{b}, \mathbf{t})$  to have  $\mathbf{v}_j = (\mathbf{v}_{bj}, \mathbf{v}_{tj})$  as the nearest neighbor. For  $k = 1$ , only the single winner fires with response value  $r_j = 1$  and all other neurons in the area don't fire  $r_i = 0$  for  $i \neq j$ .

### Learning

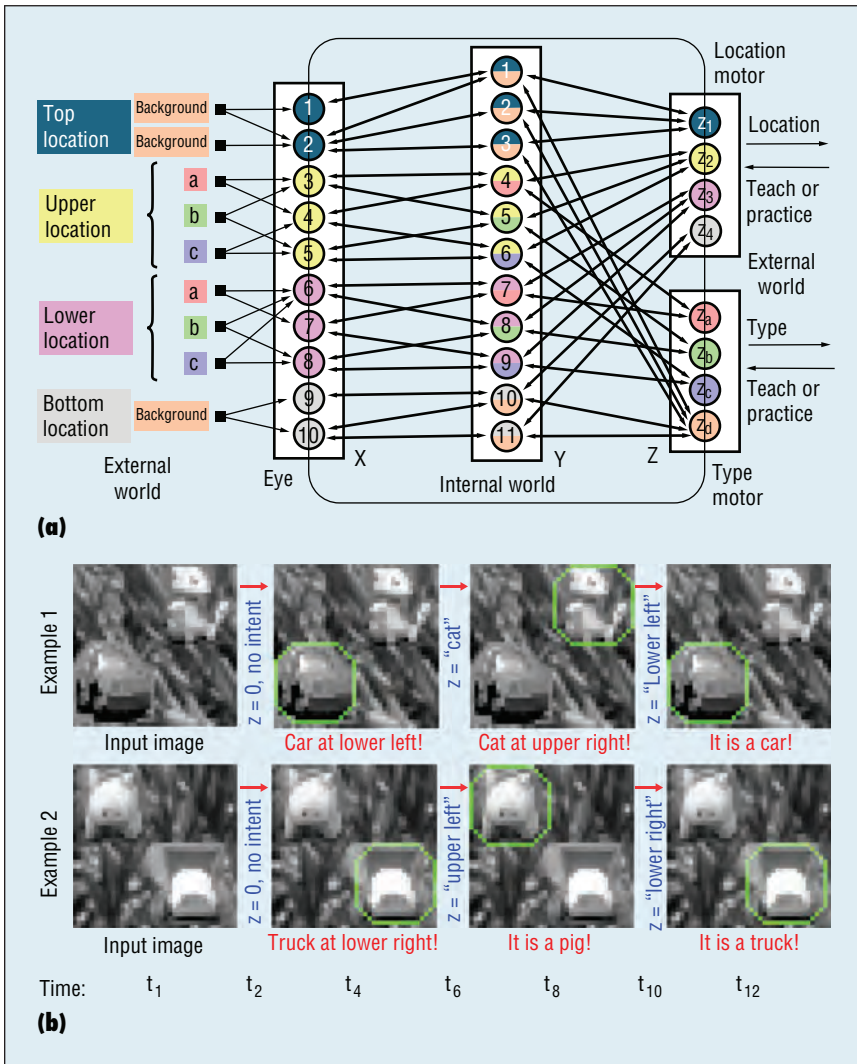
From birth, every area  $A$  of the concept network learns incrementally in the best way. With  $\hat{\mathbf{p}}(V)$ , the area's representation of input  $\mathbf{p} = (\mathbf{b}, \mathbf{t})$  based on the set of synaptic vectors  $V$ , the expected error for representing  $\mathbf{p}$  is  $E\|\hat{\mathbf{p}}(V) - \mathbf{p}\|$ . In related work, we established that the best  $V$  that minimizes the expected representation error  $E\|\hat{\mathbf{p}}(V) - \mathbf{p}\|$  is the set of lobe components  $V^*$ .<sup>15</sup> However,  $V^*$  requires infinitely many training samples. With limited experience from time  $t_0$  to time  $t_n$ , the incrementally updated set  $V(t_n)$  best updates, using the current input  $\mathbf{p}(t_n)$  and the last set  $V(t_{n-1})$  so that its distance to  $V^*$ ,  $E\|V(t_n) - V^*\|$ , is minimized (under some regularity conditions)

$$\mathbf{v}_j \leftarrow (1 - \rho(n_j))\mathbf{v}_j + \rho(n_j)r_j \mathbf{p}, \quad (4)$$

where  $\rho(n_j) = (1 + \mu)/n_j$  is the schedule of optimal learning rates at every firing age  $n_j$ , and  $r_j = 1$  for  $k = 1$ . (Note the amnesic parameter<sup>15</sup>  $\mu$ —for example,  $\mu = 2$ , is for  $\mathbf{v}_j$  to gradually disregard its estimation trajectory at an early experience.) The firing age of neuron  $j$  is incremented  $n_j \leftarrow n_j + 1$ . All nonfiring neurons don't modify their synapses nor advance their firing ages. Every area,  $X$ ,  $Y$ , and  $Z$ , computes and learns in this unified way. In other words, every area of the concept network is the quickest learner from its incremental learning experience. For an explanation on how  $V(t_n)$  is similar to and different from traditional self-organizing maps and why it's optimally described here, see our related work.<sup>15</sup>

Such living experience arises from environmental training or autonomous practice. Supervision from the environment (for example, a teacher) can be occasional. During teacher supervision, the teacher supplies a desired action  $\mathbf{z}$  (for example, pronounce "cat") for the corresponding image  $\mathbf{x}$  when the network child is staring at the object in a complex background (that is, with a correct  $\mathbf{y}_a$ ). Otherwise, the network autonomously practices by self-generating  $\mathbf{z}$  and using it for self-teaching, while  $\mathbf{y}_a$  has all 1's.

The gain vector  $r_j \mathbf{p}$  mentioned previously determines dynamic wiring. Only when the postsynaptic neuron  $j$  and presynaptic neuron  $i$  cofire does the  $i$ th component in  $\mathbf{y}_j \mathbf{p}$  become positive. As long as this component is positive once, the synapse from neuron  $i$  to the postsynaptic neuron  $j$  becomes positive. Each synapse is the incrementally estimated probability for the presynaptic neuron to cofire; conditioned on that the postsynaptic neuron fires. Therefore, when the



**Figure 2. Schematic description of concept networks. (a) A concept network. The two-color illustrations for each neuron in Y indicates that each feature neuron is a mixture of the corresponding sensory information (upper color) and motor information (lower color). Lateral connections aren't shown for simplicity. (b) Concept-free and concept-directed perception by a larger concept network. Each concept (blue) clamps (imposes) the Z area. The behavior outputs (red) are the corresponding concept-directed perception result. For fully self-generated concepts, see the homeostatic mode in Figures 3 and 4.**

postsynaptic neuron fires (that is, it's the best-matched neuron), the more often the presynaptic neuron fires, the stronger the synapse.

In addition, excitatory lateral connections and a larger  $k$  are good for earlier ages so that the weight vectors smoothly distribute across the typically lower dimensional manifolds in which the input sample  $p$ 's lie.

The following mechanic algorithm incrementally solves the task-nonspecific

learning problem of the highly non-linear, highly recurrent concept network:

1. Every area initializes its adaptive part:  $N = (V, A)$ .
2. Do the following two steps repeatedly forever while interacting with the external environment:
  - a. Every area  $A \in \{X, Y, Z\}$  computes area function Equation 1, where  $\mathbf{b}$ ,  $\mathbf{r}$ ,  $\mathbf{t}$  indicates

the bottom-up input, its own response, and the top-down input, respectively.  $r_a$  is the attention supervision vector for the Y area used only during learning, but not present for the Z area. b. Every area replaces:  $N \leftarrow N'$  and  $\mathbf{r} \leftarrow \mathbf{r}'$ .

### Experiments

This is a challenging open vision problem because every input image, except some learned unknown object patch somewhere in it, is always globally new, regardless of whether the current session is for training or testing.

Each Y neuron in our experiment has a fixed and local sensory receptive field in X, as shown in Figure 1. Other than that, every neuron connects with all neurons in its bottom-up area and top-down area (if it exists), and the automatic Hebbian learning of each synapse dynamically determines to which other neurons the neuron connects (a zero weight means no connection). The X area directly takes pixel values as neuronal responses.

The example of a trained network in Figure 2a helps our understanding. A concept network requires many neurons in Y so that any possible foreground matches at least some neurons' receptive fields. Biologically, cell migration, dendrite growth, and axonal guidance, all activity-dependent, enable each neuron to have a default input field in X and Z, respectively. Every default input field (for example, the area of the upper location in Figure 2a) is further fine-tuned through the concept network algorithm so some connections are strengthened and others are weakened or cut off. Many neurons are needed for receptive fields of many different scales at many different locations.

To train our example network, the teacher has designed a simple

language for communication with the network. Independent of network programming, the teacher has two concepts in mind, location and type (LM and TM), one for each motor subarea. She chooses the upper four motor neurons in  $Z$  in Figure 2a to represent the four values of the first concept location, and the lower four motor neurons in  $Z$  to represent those of the second concept type (see the matched colors in the figure). For example, when object  $b$  is present at the upper location, the neurons  $z_2$  and  $z_b$  are supervised to fire. All  $Y$  neurons are suppressed by supervised attention  $y_a$  except the neurons 4, 5, and 6, which compete. Suppose that the  $Y$  neuron 5 wins. Thus, its connections with  $z_2$  and  $z_b$  in  $Z$  are established because these two motor neurons are imposed to cofire. All the wires are established this way.

When the motor area isn't supervised, the network autonomously practices through self-supervision. If  $Y$  neuron 5 fires, the input must contain type  $b$  at the upper location. This  $Y$  neuron only excites the connected motor neuron  $z_2$  to report the location and the motor neuron  $z_b$  to report the type. It also boosts the corresponding  $X$  neurons 3 and 5 as top-down attention from  $Y$  to  $X$ , enabling a more steady perception. The firing of the two motor neurons  $z_2$  and  $z_b$  also boosts all the connected  $Y$  neurons as top-down attention from  $Z$  to  $Y$ . Thus, if  $Y$  has enough neurons, the network performance is perfect, as seen in Figure 2a.

We can see that the concept network learns two-way signal processing: bottom-up from sensory  $X$  to motoric  $Z$  through  $Y$ ; and top-down from motoric  $Z$  to sensory  $X$  also through  $Y$ .

For experiments of the network on video objects, stereo perception, text processing, and natural language

processing, see the review of our experiments published elsewhere.<sup>12</sup>

### Free Viewing: Performing Detection and Recognition Together

During free viewing, all  $Z$  neurons don't fire initially. In Figure 2a, suppose objects  $a$  and  $b$  are at upper and lower locations, respectively. Suppose the top  $Y$  neurons 4 and 8 have almost the same preaction potential to win:  $r_4 \approx r_8$ . Either one of the two neurons pops up to win in  $Y$ . Supposing  $r_4$  wins, the network pays attention to the upper  $a$ . The network outputs its location at  $z_2$  and type at  $z_a$ . For the examples in Figure 2b, this mode is at  $t_1$  to  $t_4$ . The information flow is  $Y \Rightarrow LM$  and  $Y \Rightarrow TM$ , as Figure 3a illustrates. If two different (or the same) objects appear on the same image but at different locations, attending either one is correct.

### Type Concept: Detection

Next, suppose a type concept cue is available, for example, an auditory signal excites motor neuron  $z_b$ , as illustrated in Figure 3b. This indicates the emergence of the type  $b$  concept from the network. Then,  $Y$  neurons 5 and 8 receive top-down boosts from  $z_b$  so the balance between  $Y$  neurons 4 and 8, mentioned previously, is broken to become  $r_4 < r_8$  and the neuron 8 wins. Then, the  $Y$  neuron 8 excites the location neuron  $z_3$  reporting the location of object  $b$ . That is, from a concept (for example, type  $b$ ), the network reasons to give another concept (for example, location). In Figure 2b, this mode is in Example 1 during  $t_6$  to  $t_8$ . The information flow is  $TM \Rightarrow Y \Rightarrow LM$ , while every area in the network updates, as Figure 3b illustrates.

### Location Concept: Recognition

Likewise, the interaction with the environment gives rise to a location

concept in  $Z$ . The network gives another concept (type). This mode corresponds to other remaining cases in Figure 2b. The information flow is  $LM \Rightarrow Y \Rightarrow TM$ , while every area in the network updates, as Figure 3c illustrates.

### Homeostatic Concept: Shift Attentions

In the homeostatic (or habituated) mode, the network self-alters its concept. The currently firing motor neuron gets suppressed, simulating, for example, temporary local depletion of neural transmitters, as Figure 3d illustrates. Another motor neuron wins. A new concept emerges as the new winner motor neuron, and the network carries out this new concept, externally shown as attention shift.

All of these modes switch automatically in a living concept network, depending on whether the environment provides a cue for concept (location, type, or any other concept learned) or not.

### Setting

More than 75 percent pixels were from unknown backgrounds when a single foreground object is present. The network area sizes, in terms of the number of neurons are as follows.

- image area  $X$ :  $38 \times 38$ ;
- internal area  $Y$ :  $20 \times 20 \times 3 = 1,200$ ; and
- motor area  $Z$ : LM area is  $20 \times 20 = 400$  locations, and TM area is  $5 \times 1$  (one for each of the five object types).

As Figure 4 shows, we used three views from each object for training, amounting to  $400 \times 3 = 1,200$  images for each object class. We used two other views for each object class for testing at all locations. The total number of foreground patches for training amounts to  $1,200 \times 5 = 6,000$ , but there are only 1,200

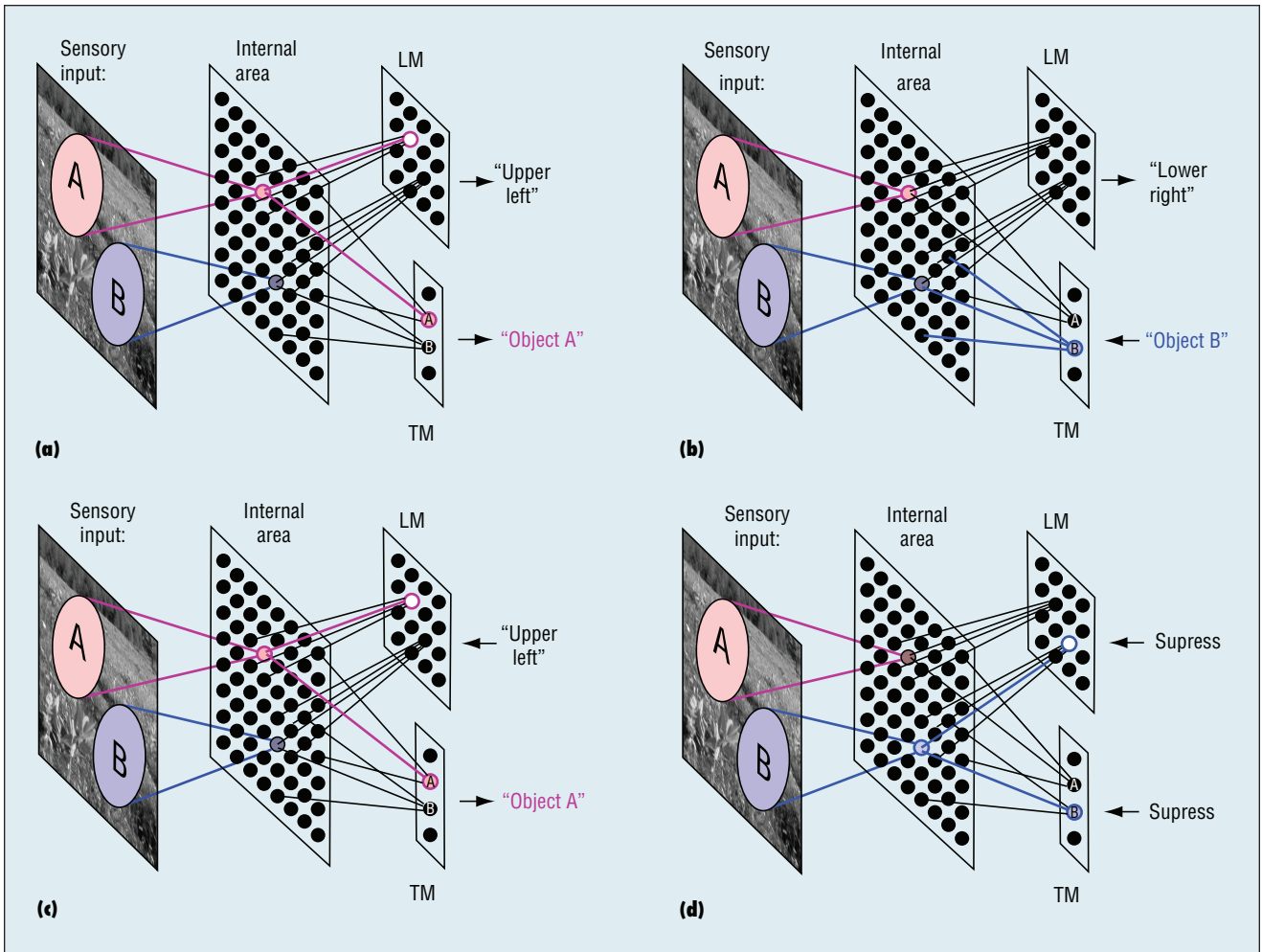


Figure 3. The concept network performs four tasks using the same network, depending on the context in the Z area. (a) Free viewing: without any Z context, perform both detection to output location and recognition to output type. (b) Object detection: from a type context in Z, perform object detection. (c) Object recognition: from a location context in Z, perform object recognition. (d) Homeostasis: suppress the current outputs from Z, and perform attention shifts through homeostasis.

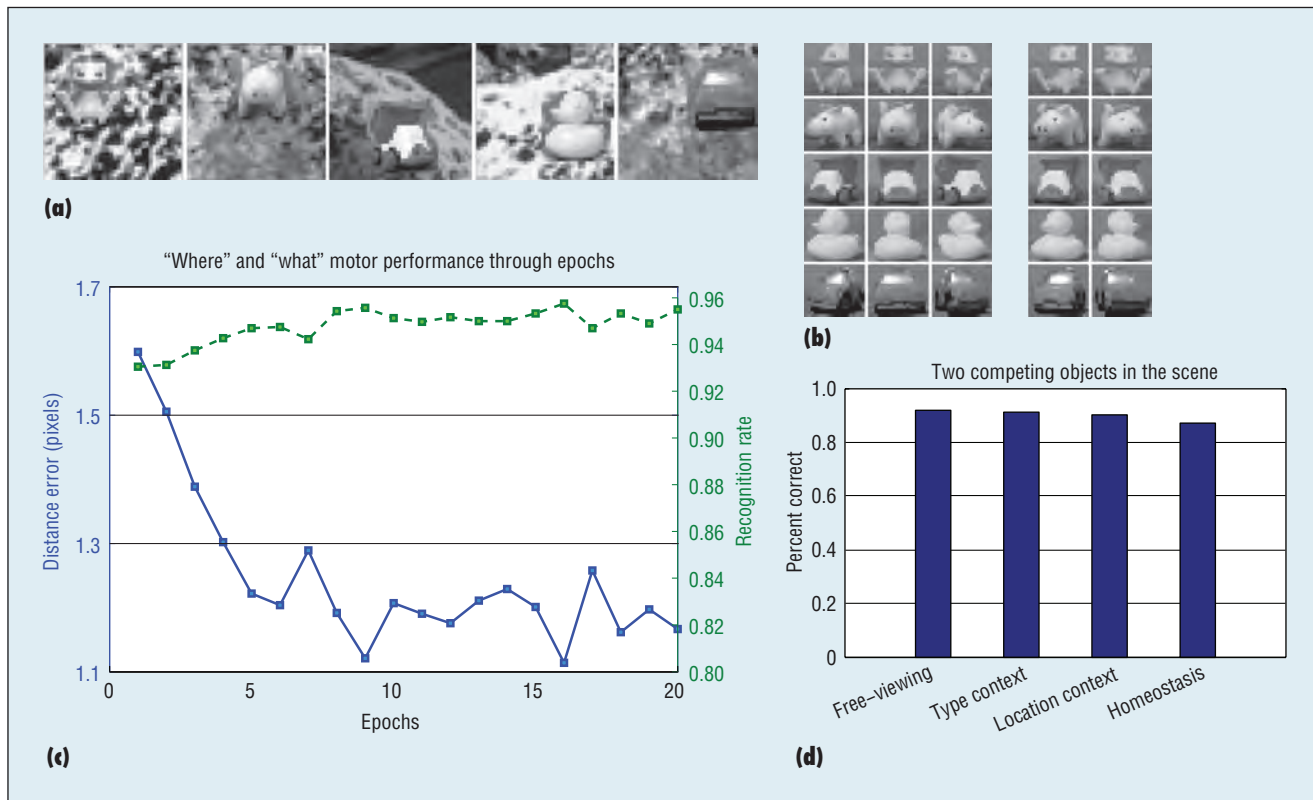
neurons in Y. This means that the Y area in this limited-size network has only  $1,200/6,000 = 20$  percent of the neurons needed to memorize all the foreground patches.

In addition, each area simulates the six-layer structure in the laminar cortex for pre-screening which reduces *top-down hallucination*, as discussed elsewhere.<sup>17</sup>

**A** neural firing pattern can represent any abstract concept, fundamentally different from any atomic symbol whose meaning is statically handcrafted by a human

(see a review about emergent representations and symbolic representations in other work<sup>11</sup>). Natural brains tell us that such firing patterns that represent abstract concepts must be open in the agent’s effector end—as both outputs and inputs instead of being hidden inside the closed brain skull—to be executable, testable, calibratable, adaptable, and enrichable with the physical world that includes human teachers. This is true regardless of whether the concepts represent declarable skills (for example, writing an object type) or non-declarable skills (for example, riding a bike).

The key of the theory and algorithm here is how to enable the concept network to develop an abstract concept that’s invariant to concrete examples (for example, small image patches appearing on an object) in infinitely many, cluttered, and new scenes. Because the network receives location and type concepts in its two motor areas as part of lifetime experience and the network programming doesn’t use any meaning of such concepts, we predict that, in principle, a concept network can learn many more practical concepts (for example, scale<sup>1</sup>) and perform concept-directed perception and action. One way is to



**Figure 4.** The experiments with a limited-size concept network. (a) Sample input images from 2,000 training images in each epoch and disjoint tests using different background images. (b) Foreground images for training (left columns) for each object and test images (right columns) from different viewing angles. (c) The average errors of actions in the free-viewing mode ( $z = 0$  originally)—detected object location and recognized object type of the attended object in each input image—in new complex natural backgrounds.<sup>16</sup> (d) Summary when input contains two learned objects in four modes: free-viewing mode (object detection and object recognition together), type-context (object detection), location context (object recognition), and fully autonomous concepts (homeostasis).<sup>16</sup>

increase the number of concept zones in  $Z$ .<sup>1</sup> Another way, like a human, is to use a more complex motor language in which a different motor sequence means a different concept and concept value. The human programmer doesn't need to know about the concepts that the network will end up learning in the future, nor its language, because the representations for learned concepts are all emergent, including all the  $X$ ,  $Y$ , and  $Z$  areas. The more the network learns, explores, and practices, hopefully, the smarter it becomes. In other work, Juyang Weng further discusses the optimality that points to the scalability of the concept networks here.<sup>18</sup>

This work seems to have shown that all feed-forward computation models are short of those basic brain functions

other than global pattern classification,<sup>19,20</sup> regardless of whether they use a spiking neuronal model or a firing rate model (this model applies to both neuronal models depending on the network update rate).

However, much engineering work is needed before a robot can learn fully autonomously as this theoretical model proposed. We subscribe to the principle of scaffolding, pioneered by Lev S. Vygotsky,<sup>21</sup> a concept that's well known in developmental psychology<sup>22</sup>—early simple skills in a life assist in the emergence of later complex skills in the life. For example, coarse location and type skills early in life assist the autonomous development of finer location skills, finer type skills, and new scale skills later in life.<sup>23</sup> This new work is principled in nature,

not for comparison with benchmarks from methods that weren't for the right three columns of Table 1 or to immediately demonstrate higher concepts (for example, jealousy). Like any new major technology that arose in the past, the participation of existing industries and the birth and growth of new industries are powerful driving forces in future applications. Such industries would produce practical concept agents and robots that artificially live in the real, physical world together with humans to autonomously develop increasingly more complex concepts and skills. ■

### Acknowledgments

Juyang Weng conceptualized and drafted this article. Matthew D. Luciw performed the experiments shown in Figures 2b and 4 when he was at Michigan State University.




## THE AUTHORS

**Juyang Weng** is a professor of computer science and engineering and a faculty member of the cognitive science program and the neuroscience program at Michigan State University. His research interests include 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. Weng has a PhD in computer science from the University of Illinois at Urbana-Champaign. He is editor in chief of the *International Journal of Humanoid Robotics* and of the *Brain-Mind Magazine*, associate editor of the *IEEE Transactions on Autonomous Mental Development*, and president of the Brain-Mind Institute. Contact him at weng@cse.msu.edu.

**Matthew D. Luciw** is a research scientist at the Center for Computational Neuroscience and Neural Technology, Boston University, and is affiliated with Neurala. His research interests include developmental robotics, neural networks, artificial curiosity, unsupervised learning, and reinforcement learning. Luciw has a PhD in computer science from Michigan State University. Contact him at luciwmat@bu.edu.

## References

1. X. Song, W. Zhang, and J. Weng, "Where-What Network 5: Dealing with Scales for Objects in Complex Backgrounds," *Proc. Int'l Joint Conf. Neural Networks*, 2011, pp. 2795–2802.
2. R. Desimone and J. Duncan, "Neural Mechanisms of Selective Visual Attention," *Ann. Rev. of Neuroscience*, vol. 18, 1995, pp. 193–222.
3. M. Corbetta, "Frontoparietal Cortical Networks for Directing Attention and the Eye to Visual Locations: Identical, Independent, or Overlapping Neural Systems?" *Proc. Nat'l Academy of Sciences*, vol. 95, no. 3, 1998, pp. 831–838.
4. L. Itti and C. Koch, "Computational Modelling of Visual Attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, 2001, pp. 194–203.
5. T.J. Buschman and E.K. Miller, "Top-Down versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices," *Science*, vol. 315, no. 5820, 2007, pp. 1860–1862.
6. L. Itti and C. Koch, "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention," *Vision Research*, vol. 40, nos. 10–12, 2000, pp. 1489–1506.
7. E.I. Knudsen, "Fundamental Components of Attention," *Ann. Rev. of Neuroscience*, vol. 30, 2007, pp. 57–78.
8. C.H. Anderson and D.C. Van Essen, "Shifter Circuits: A Computational Strategy for Dynamic Aspects of Visual Processing," *Proc. Nat'l Academy of Sciences*, vol. 84, no. 17, 1987, pp. 6297–6301.
9. 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," *J. Neuroscience*, vol. 13, no. 11, 1993, pp. 4700–4719.
10. J.K. Tsotsos et al., "Modeling Visual Attention via Selective Tuning," *Artificial Intelligence*, vol. 78, nos. 1–2, 1995, pp. 507–545.
11. J. Weng, "Symbolic Models and Emergent Models: A Review," *IEEE Trans. Autonomous Mental Development*, vol. 4, no. 1, 2012, pp. 29–53.
12. J. Weng, "A 5-Chunk Developmental Brain-Mind Network Model for Multiple Events in Complex Backgrounds," *Proc. Int'l Joint Conf. Neural Networks*, 2010, pp. 1–8.
13. D.J. Felleman and D.C. Van Essen, "Distributed Hierarchical Processing in the Primate Cerebral Cortex," *Cerebral Cortex*, vol. 1, no. 1, 1991, pp. 1–47.
14. G.E. Hinton, "Learning Multiple Layers of Representation," *Trends in Cognitive Science*, vol. 11, no. 10, 2007, pp. 428–434.
15. J. Weng and M. Luciw, "Dually Optimal Neuronal Layers: Lobe Component Analysis," *IEEE Trans. Autonomous Mental Development*, vol. 1, no. 1, 2009, pp. 68–85.
16. J. Weng and M. Luciw, "Brain-Like Emergent Spatial Processing," *IEEE Trans. Autonomous Mental Development*, vol. 4, no. 2, 2012, pp. 161–185.
17. Z. Ji and J. Weng, "WWN-2: A Biologically Inspired Neural Network for Concurrent Visual Attention and Recognition," *Proc. IEEE Int'l Joint Conf. Neural Networks*, 2010, pp. 1–8.
18. J. Weng, "Why Have We Passed 'Neural Networks Do Not Abstract Well'?" *Natural Intelligence: The INNS Magazine*, vol. 1, no. 1, 2011, pp. 13–22.
19. Y. LeCun et al., "Gradient-Based Learning Applied to Document Recognition," *Proc. IEEE*, vol. 86, no. 11, 1998, pp. 2278–2324.
20. Q. Yu et al., "A Brain-Inspired Spiking Neural Network Model with Temporal Encoding and Learning," *Neurocomputing*, vol. 138, 2014, pp. 3–13.
21. L.S. Vygotsky, *Thought and Language*, MIT Press, 1962.
22. D.J. Wood, J.S. Bruner, and G. Ross, "The Role of Tutoring in Problem Solving," *J. Child Psychology and Psychiatry*, vol. 17, no. 2, 1976, pp. 89–100.
23. Z. Zheng et al., "WWN: Integration with Coarse-to-Fine, Supervised and Reinforcement Learning," *Proc. Int'l Joint Conf. Neural Networks*, 2014, pp. 1–8.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.