Where-What Network 1: "Where" and "What" Assist Each Other Through Top-down Connections

Zhengping Ji and Juyang Weng
Embodied Intelligence Laboratory
Michigan State University, Lansing, USA
Email: {jizhengp, weng}@cse.msu.edu

Danil Prokhorov
Toyota Technical Center
Ann Arbor, USA
Email: danil.prokhorov@tema.toyota.com

***Abstract*—This paper describes the design of a single learning network that integrates both object location ("where") and object type ("what"), from images of learned objects in natural complex backgrounds. The in-place learning algorithm is used to develop the internal representation (including synaptic bottom-up and top-down weights of every neuron) in the network, such that every neuron is responsible for the learning of its own signal processing characteristics within its connected network environment, through interactions with other neurons in the same layer. In contrast with the previous fully connected MILN [13], the cells in each layer are locally connected in the network. Local analysis is achieved through multi-scale receptive fields, with increasing sizes of perception from earlier to later layers. The results of the experiments showed how one type of information ("where" or "what") assists the network to suppress irrelevant information from background (from "where") or irrelevant object information (from "what"), so as to give the required missing information ("where" or "what") in the motor output.**

I. Introduction

The primate visual pathways have been extensively investigated in neuroscience: branching primarily from V2, two primary pathways exist, called the dorsal pathway and the ventral pathway, respectively. The dorsal stream begins with V1, through V2, the dorsomedial area and MT (also known as V5), to the posterior parietal cortex. The control of attention employment is believed to mostly take place in the dorsal pathway, sometimes called the "where" pathway. The ventral stream begins from V1, through V2, V4, and to the inferior temporal cortex. The ventral stream, also called the "what" pathway, is mainly associated with the recognition and identification of visual stimuli.

Attention and recognition are known as a chicken-and-egg problem. Without attention, recognition cannot do well; recognition requires attended areas for the further processing. Without recognition, attention is limited; attention does not only need bottom-up saliency-based cues, but also top-down target-dependant signals.

Bottom-up Attention Studies in psychology, physiology, and neuroscience provided qualitative models for the bottom-up attention, i.e., attention uses different properties of sensory inputs, e.g., color, shape, and illuminance to extract saliency. The first explicit computational model of bottom-up attention was proposed by Koch & Ullman in 1985 [6], in which a "saliency map" is used to encode stimuli saliency at every lactation in the visual scene. More recently, Itti & Koch et al. 1998 [5] integrated color, intensity, and orientation as basic features, and extracted intensity information in six scales for attention control. An active-vision system, called NAVIS (Neural Active Vision) by Baker et al. 2001, was proposed to conduct the visual attention selection in a dynamic visual scene [1].

Top-down Attention Volitional shifts of attention are also thought to be performed top-down, through spacial defined and feature-dependant weighting of the various feature maps. The successful modeling of the "where" pathway, then, involves the integration of bottom-up and top-down cues, such as to provide coherent control signals for the focus of attention, and the interplay between attentional tuning and object recognition. Olshausen et al. 1993 [8] proposed a model of how visual attention can solve the object-recognition problem of position and scale invariance. A top-down attention model was discussed by Tsotsos et al. 1995 [12], who implemented attention selection using a combination of a bottom-up feature extraction scheme and a top-down selective tuning scheme. Mozer et al. 1996 proposed a model called MORSEL [7], to combine the object recognition and attention, in which the attention is shown to help recognition. Rao et al. 2004 [9] described an approach that allowed a pair of cooperating neural networks, to estimate object identity and object transformations, respectively. A top-down, knowledge-based recognition component, presented by a hierarchical knowledge tree, was introduced by Schill et al. 2001 [10], where object classes were defined by several critical points and the corresponding eye movement commands that maximize the information gain. Deco & Rolls 2004 [2] presented a model of invariant object recognition that incorporated feedback biasing of top-down attentional mechanisms on a hierarchically organized set of visual cortical areas. A more extreme view is expressed by the "scanpath theory" of Stark & Choi 1996 [11], in which the control of eye movements is almost exclusively under top-down control.

Aforementioned mechanisms of selective visual attention play significant roles in the biologically plausible architectures for object recognition (called **attention-based recognition**) in the ventral " stream". However, it remains an open issue for the recognition models to integrate neurobiological models concerned with attentional control in the dorsal "where" stream. As pointed out by Itti & Koch 2001 [4], this integration will,

in particular, account for the increasing efforts on an **object-based spatial attention**.

In this paper, we propose a developmental network, called "Where-What" Network 1 (WWN-1), for a general sensorimotor pathway, such that recognition and attention interact with each other in a single network. As this is a very challenging design and understanding task, we concentrate on (1) the network design issue: how such a network can be designed so that attention and recognition can assist each other; (2) how to understand a series of theoretical, conceptual, and algorithmic issues that arise from such a network. To verify the mechanisms that are required for both design and understanding, in the results presented, we limit the complexity of "where" and "what" outputs,

The following technical characteristics required by developmental learning make such work challenging: (1) Integrate both bottom-up and top-down attention; (2) Integrate attention-based recognition and object-based spacial attention interactively; (3) Enable supervised and unsupervised learning in any order suited for development; (4) Local-to-global invariance from early to later processing, through multi-scale receptive fields; (5) In-place learning: each neuron adapts "in-place" through interactions with its environment and it does not need an extra dedicated learner (e.g., to compute partial derivatives). The WWN-1 uses the top-k mechanism to simulate in-place competition among neurons in the same layer, which is not in-place per se but is still local and computationally more efficient as it avoids iterations with a layer. Rather than the simulations of fMRI data, the engineering performance of recognition rate and attended spatial locations are presented for an image dataset in the experiment.

In what follows, we first explain the structure of the proposed WWN-1. Key components of the model are presented in Section III, IV, V, addressing local receptive field, cortical activation and lateral connections, respectively. Section VI provides the algorithm of weight adaptation in the proposed network. Experimental results are reported in Sec. VII and concluding remarks are provided in Sec. VIII.

II. Network Overview

Structurally, the "Where-What" Network 1 is a set of connected two-dimensional cortical layers, each containing a set of neurons, arranged in a multi-level hierarchy. The number of levels of neurons is determined by the size of local receptive fields and staggered distance, discussed in Sec. III. An example of the network is shown in Fig.1. Its network architecture and parameters will be used in our experiments of Sec. VII. The network operates at discrete times $t = 0, 1, ...$ Each neuron is placed at a 2D position in a layer, so that each layer forms a grid of $n \times n$ neurons.

The external sensors are considered to be on the bottom (layer 0) and the external motors on the top (layer N). Neurons are interconnected with nonnegative weights. For each neuron (i, j), at level l $(0 < l < N)$, there are four weight vectors, as illustrated in Fig.2:

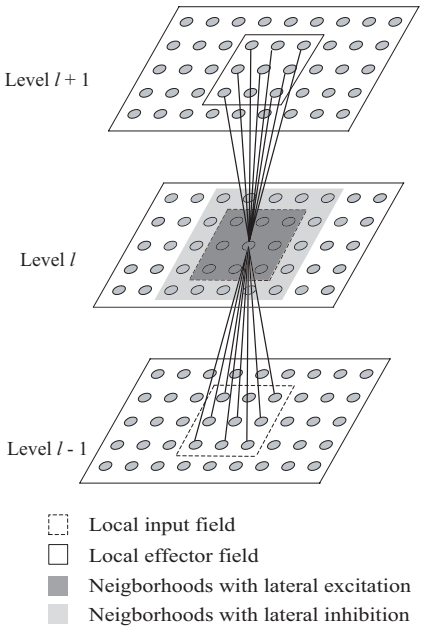


Fig. 2. For in-place learning, neurons are placed (given a position) on different levels in an end-to-end hierarchy – from sensors to motors. A neuron has feed-forward, horizontal, or feedback projections to it.

1) bottom-up weight vector $\mathbf{w^b}_{i,j}$ that links connections from its local input field in the previous level;
2) top-down weight vector $\mathbf{w^t}_{i,j}$ that links connections from its effector field, either local or global, in the next level;
3) lateral weight vector $\mathbf{w^h}_{i,j}$ that links inhibitory connections from neurons in the same layer (long range).
4) lateral weight vector $\mathbf{w^e}_{i,j}$ that links excitatory connections from neurons in the same layer (short range).

III. Local Receptive Fields

Hubel and Wiesel (e.g., [3]) explained that receptive fields of cells at one cortical area of the visual system are determined by input from cells at an earlier area of the visual system. In this manner, small, simple receptive fields could be combined to form large, complex receptive fields. Localized connections are utilized in the WWN-1, providing a structural basis for local attention. Attention selection needs to suppress neuronal responses whose receptive fields fall out of the attended receptive field.

Each neuron receives its input from a restricted region in the previous layer, called local input field. Fig. 3 shows the organization of square input fields in a layer consisting of $n \times n$ neural units. Where a connection falls outside of the neuronal plane, the input is always 0. Let S_l and d_l be the number of neurons and staggered distance in the current layer l. The total number of input fields, namely, the number of neurons in the next layer is thus determined by:

$$S_{l+1} = \left(\frac{\sqrt{S_l}}{d_l} \right)^2 \tag{1}$$

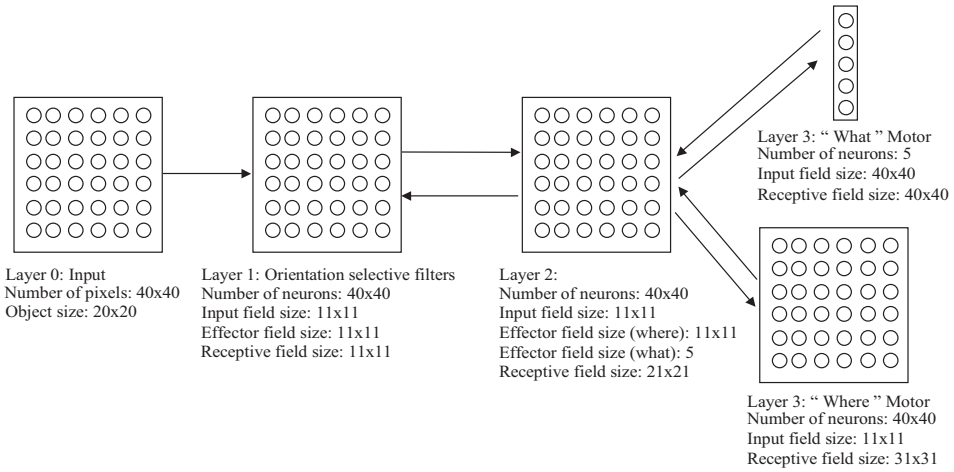


Fig. 1. The specific network architecture and parameters in our experiment.

For $n \times n$ neurons shown in Fig. 3, therefore, there are $n \times n$ neurons in the next layer, when the staggered distance is set to be 1.

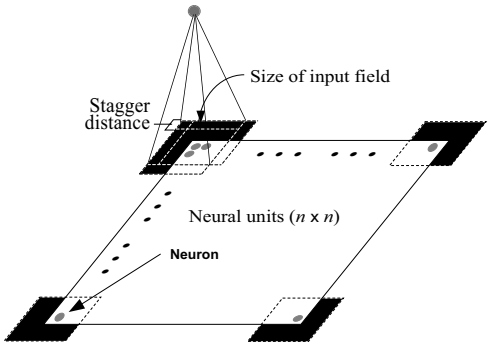


Fig. 3. Input field boundaries and numbering scheme for neurons in a layer. When the local input field falls out of the input neuronal plane, the corresponding inputs are zeros (black areas in the figure).

The overlapped square input fields allow the network to obtain alternative receptive fields at multiple scales and positions. Fig. 4 shows how the receptive fields increase from one layer to the next until the entire input is covered with a single receptive field. This representation provides information for receptive fields at different locations and with different sizes.

IV. Cortical Activation

From Layer 1 to Layer $N-2$ of the proposed network, the layer responses are computed the same way as described in [13], except that local connections are applied here. The pre-response $z_{i,j}$ of the neuron (i,j) is determined by

$$z_{i,j} = g_i\left((1-\alpha)\frac{\mathbf{w^b}_{i,j}(t)\cdot\mathbf{x}_{i,j}(t)}{\|\mathbf{w^b}_{i,j}(t)\|\|\mathbf{x}_{i,j}(t)\|}+\alpha\frac{\mathbf{w^t}_{i,j}(t)\cdot\mathbf{y}_{i,j}(t)}{\|\mathbf{w^t}_{i,j}(t)\|\|\mathbf{y}_{i,j}(t)\|}\right)$$

where $\mathbf{x}_{i,j}$ is the local bottom-up input and $\mathbf{y}_{i,j}$ is the local or global top-down input. g is its nonlinear (or a piecewise linear approximation) sigmoidal function. α $(0 \le \alpha \le 1)$ denotes a specific weight that controls the maximum contribution by the top-down versus the bottom-up. The length normalization of $\mathbf{x}_{i,j}$ and $\mathbf{y}_{i,j}$ ensures that the bottom-up part and top-down part are equally scaled.

Layer $N-1$, however, receives the top-down projection from both "where" motor layer and "what" motor layer (see Fig. 5). Thus, the pre-response $z_{i,j}$ of the neuron (i,j) is

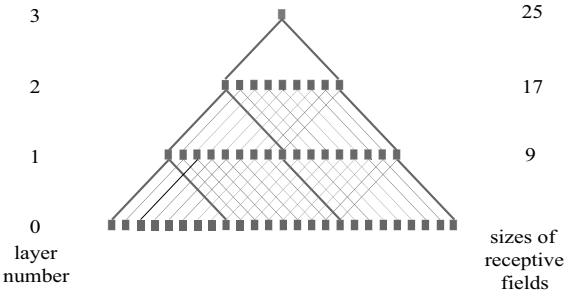


Fig. 4. The architecture of receptive fields in different scales and positions. The size of the receptive field in a particular layer is 8 larger than its previous layer in this diagram (shown at the right), whereas the size of input field is set to be 9 at each layer.

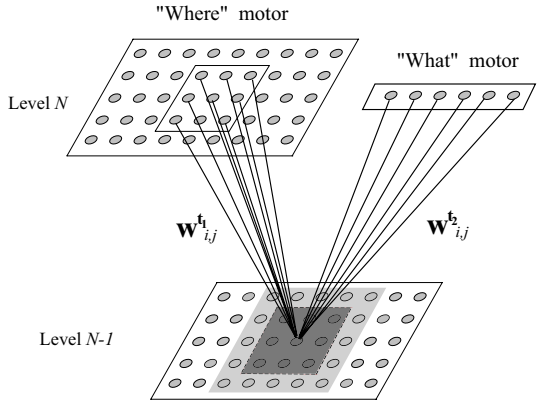


Fig. 5. Top-down projection onto Layer $N-1$.

determined by

$$z_{i,j} = g_i\Bigg((1-\alpha)\frac{\mathbf{w^b}_{i,j}(t)\cdot\mathbf{x}_{i,j}(t)}{\|\mathbf{w^b}_{i,j}(t)\|\|\mathbf{x}_{i,j}(t)\|}+$$

$$\alpha\big((1-\beta)\frac{\mathbf{w^{t_1}}_{i,j}(t)\cdot\mathbf{y^1}_{i,j}(t)}{\|\mathbf{w^{t_1}}_{i,j}(t)\|\|\mathbf{y^1}_{i,j}(t)\|}+\beta\frac{\mathbf{w^{t_2}}_{i,j}(t)\cdot\mathbf{y^2}_{i,j}(t)}{\|\mathbf{w^{t_2}}_{i,j}(t)\|\|\mathbf{y^2}_{i,j}(t)\|}\big)\Bigg)$$

$\mathbf{w^{t_1}}_{i,j}$ and $\mathbf{w^{t_2}}_{i,j}$ are top-down weights received from "where" and "what" motors, respectively. $\mathbf{y^1}_{i,j}$ and $\mathbf{y^2}_{i,j}$ are the top-down inputs from the "where" motor and "what" motor, respectively. β $(0 \le \beta \le 1)$ is the weight that controls the maximum contribution by the "what" motor.

V. Neuron Competition

Lateral inhibition is a mechanism of competition among neurons in the same layer. The output of neuron A is used to inhibit the output of neuron B, which shares a part of the input field with A, totally or partially. As an example shown in Fig.6, the neighborhood of lateral inhibition contains $(2h-1)\times(2h-1)$ neurons, because neuron (i,j) and $(i,j-h)$ do not share any input field at all. We realize that the net effect of lateral inhibition is (a) for the strongly responding neurons to effectively suppress weakly responding neurons, and (b) for the weakly responding neurons to less effectively suppress strongly responding neurons. Since each neuron needs the output of other neurons in the same layer and they also need the output from this neuron, a direct computation will require iterations, which is time consuming. To avoid iterations, we use the following local top-k mechanism.

For any neuron (i,j) in the layer l $(1 < l \le N)$, we sort all the pre-responses from neurons, centered at positions inside the input field of neuron (i,j). After sorting, they are in order: $z_1 \ge z_2 \ge ... \ge z_m$. The pre-responses of top-k responding neurons are scaled with non-zero factor. All other neurons in the neighborhood have zero responses. Suppose the pre-response $z_{i,j}$ of neuron (i,j) is the top q in the local inhibitory neighbors, i.e. $z_{i,j} = z_q$. Its response $z'_{i,j}$ is then

$$z'_{i,j} = \begin{cases} z_{i,j}\times(z_q - z_{k+1})/(z_1 - z_{k+1}) & \text{if } 1 \le q \le k \\ 0 & \text{otherwise} \end{cases}$$

In other words, if the pre-response of neuron (i,j) is the local top-1, then this response is the same as its pre-response. Otherwise, its pre-response is lower than its pre-response, to simulate lateral inhibition. A larger k gives more information about the position of the input in relation with the top-k winning neurons. However, an overly large k will violate the sparse coding principle (i.e., neurons should be selective in responding so that different neurons detect different features). In our experiments, k is set at 5% of the number of neurons in the local input field. Sparse-coding is a result of lateral inhibition, stimulated by the local top-k rule. It allows relatively few winning neurons to fire in order to disregard less relevant feature detectors.

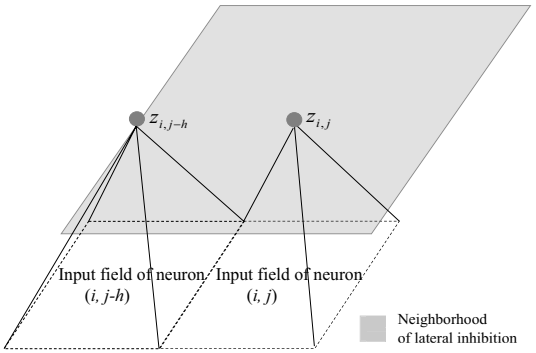


Fig. 6. The neuron (i,j) has a $(2h-1)\times(2h-1)$ neighborhood of lateral inhibition, while neuron (i,j) and neuron $(i,j-h)$ did not share any input fields.

VI. Weight Adaptation

After the responses have been computed, the connection weights of each neuron are updated if the neuron has non-zero response. Both the bottom-up and top-down weights adapt according to the same biologically motivated mechanism: the Hebb rule. For a neuron (i,j) with non-zero response (along with its 3×3 neighboring neurons), the weights are updated using the neuron's own internal temporally scheduled plasticity:

$$\begin{cases} \mathbf{w^b}_{i,j}(t) = \omega_1\mathbf{w^b}_{i,j}(t-1) + \omega_2 z'_{i,j}\mathbf{x}_{i,j}(t) \\ \mathbf{w^t}_{i,j}(t) = \omega_1\mathbf{w^t}_{i,j}(t-1) + \omega_2 z'_{i,j}\mathbf{y}_{i,j}(t) \end{cases}$$

The 3×3 updating rule is to model the lateral excitation on the short-range neighboring neurons, in order to achieve a smooth representation across the layer.

The scheduled plasticity is determined by its two age-dependent weights:

$$\omega_1 = \frac{n_{i,j} - 1 - \mu(n_{i,j})}{n_{i,j}}, \omega_2 = \frac{1+\mu(n_{i,j})}{n_{i,j}},$$

where $n_{i,j}$ is the number of updates that the neuron has gone through, with $\omega_1 + \omega_2 \equiv 1$. $\mu(n_{i,j})$ is a plasticity function defined as

$$\mu(n_{i,j}) = \begin{cases} 0 & \text{if } n_{i,j} \le t_1, \\ c\times(n_{i,j}-t_1)/(t_2-t_1) & \text{if } t_1 < n_{i,j} \le t_2, \\ c + (n_{i,j}-t_2)/r & \text{if } t_2 < t \end{cases}$$

where plasticity parameters $t_1 = 20, t_2 = 200, c = 2, r = 2000$ in our implementation.

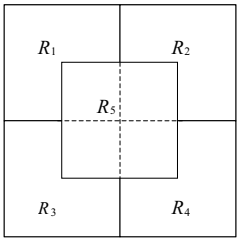
Finally, the neuron age $n_{i,j}$ is incremented: $n_{i,j} \leftarrow n_{i,j}+1$. All neurons that do not fire (i.e., zero-response neurons) keep their weight vector and age unchanged for long-term memory.

VII. Experiment

Fig. 1 shows a specific set-up of parameters and architecture implemented in our experiment, where $\alpha = 0.3$, $\beta = 0.5$ for the training process. As a first study of the proposed framework, "what" motors are simplified to define 5 different objects, which are shown in Fig. 7(a). The images of objects are normalized in size, in this case to 20 rows and 20 columns.

Each object is placed in 5 different regions ($R_1, ..., R_5$ in Fig 7(b)), defined by "where" motors. For each object-position combination, different backgrounds (each has 40×40 dimensions) are randomly selected from natural images[1]. Thus, there were 5 (positions) $\times 20$ (backgrounds)$=$ 100 samples for each object class, and 500 samples in total.

(a)



(b)

Fig. 7. (a) Five objects defined by "what" motor. When an object, e.g. "apple" appears in the image, the corresponding neuron in "what" motor is set to be 1 and all others to be 0. (b) Five regions defined by "where" motors. When an object appears in one region, e.g., R_1, all the neurons in R_1 are set to be 1 and others set to be 0. The "what" and "where" motors supervise the learning of neurons weights in previous layers, through the top-down connections described in Fig. 5

A. Development of Layer 1

We first develop the features in Layer 1 of the proposed model. 500,000 of 40×40-pixel image patches are randomly selected from thirteen natural images [1] (no object presence), learnt through the in-place learning algorithm described from Sec. III to Sec. VI, without supervision by motors ($\alpha = 1$ by the off-line feature development). The developed bottom-up synaptic weights of all neurons in Layer 1 are shown as image patches in Fig. 8. They clearly show localized patterns because each has a local input field with size 11×11. Many of the developed features resemble the orientation selective cells.

B. Recognition Through Attention

To evaluate the performance of recognition, the network weights are incrementally updated using one frame of training images at a time. After the network updated for each training sample, the network is tested for the recognition rate of all the samples, where $\beta = 0$ to disable the top-down supervision from "what" pathway. The attended region is supervised by the "where" motor, using 11×11 local effector fields, to guide the agent's attention. As shown in Fig. 9, with the guided attention, approximately 25% of samples are sufficient to achieve a 90% recognition rate. However, the recognition rate is only about

[1]available from http://www.cis.hut.fi/projects/ica/imageica/

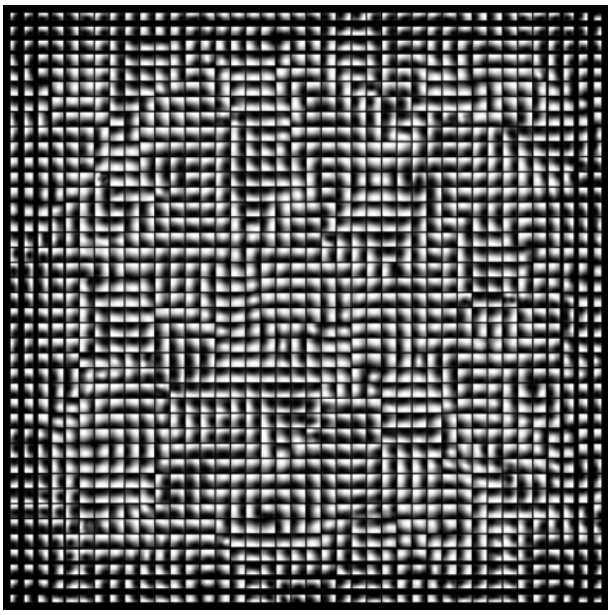


Fig. 8. Bottom-up synaptic weights of neurons in Layer 1, developed through randomly selected patches from natural images.

45% if the attention motor is not available (all zeros) during the testing. This is a test to show how top-down "where" supervision helps recognition of "what".

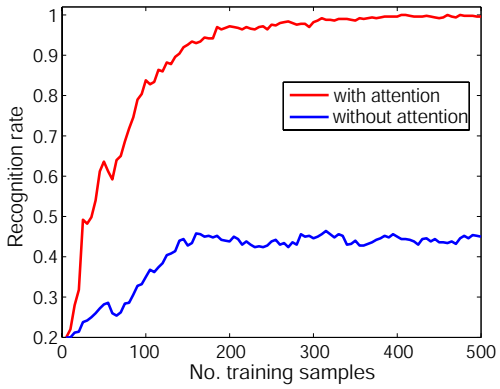


Fig. 9. Recognition rate with incremental learning, using one frame of training images at a time.

C. Attention Through Recognition

To examine the effect of top-down "what" supervision in identifying where the object is, we only supply the information of "what" in the "what" motor during tests, where $\beta = 1$. The representation of supervision here is global , i.e., the input size of top-down connection from "what" motor is 5. Examples of attention results are shown in Fig.10, where the network presents better attention capability with the assistance of "what" supervision. This is a test for how top-down "what" supervision helps location finding of "where".

The bottom-up weights of "what" and "where" motors are shown in Fig. 11. The Fig. 11(a) shows that each "what"

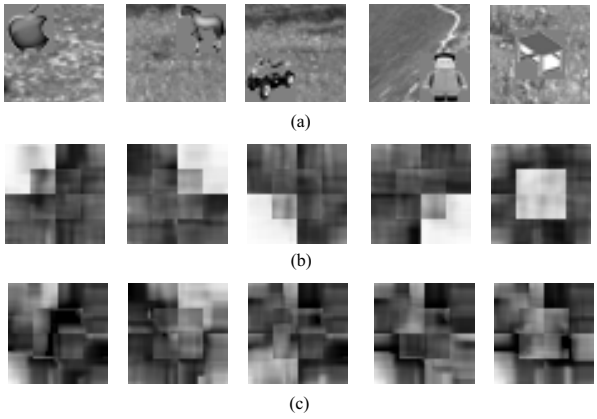


Fig. 10. (a) Examples of input images; (b) Responses of attention ("where") motors when supervised by "what" motors. (c) Responses of attention ("where") motor when "what" supervision is not available.

motor detects the corresponding features at different locations (i.e., position invariance for "what" motors). The Fig. 11(b) indicates that each "where" motor's bottom-up weight vector gives the average pattern in its input field across different objects. They are selective as not every input component fires across different objects.

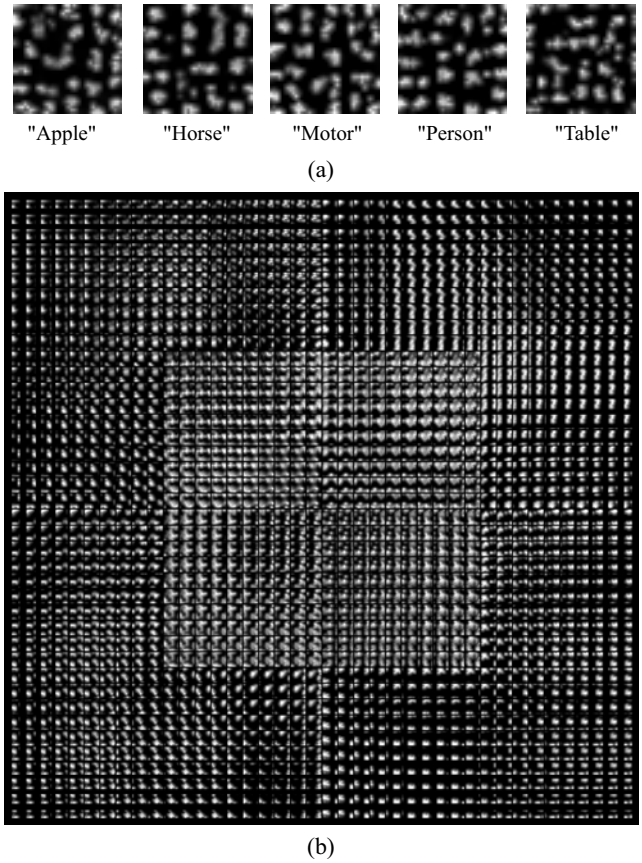


Fig. 11. (a) Bottom-up weights of "what" motors in Layer 3. (b) Bottom-up weights of "where" motors in Layer 3.

VIII. Conclusion

Locally connected WWN-1 proposes local feature detectors at every layer. When two kinds of motor layers are connected with Layer 2, top-down connections from one motor layer helps the output from another motor layer in an interesting way. Specifically, (1) when the "what" motor is on during stimuli presentation, the features that are (learned to be) associated with this particular object are boosted from top-down attention (i.e., expectation). These boosted object-specific features suppress the features that respond to background. Such suppression enables the "where" motors to report locations where features are boosted. (2) Conversely, when the "where" motor is on during stimuli presentation, the features that are (learned to be) associated with this "where" motor are boosted from top-down attention (i.e., covert attention instead of overt eye saccade). These boosted features corresponding to attended object suppress the features that respond to background. Such suppression leads to a significant boost in foreground recognition rate with presented natural background (from 45% to 100% in the experiment). Both the bottom-up and top-down attention mechanisms have been integrated in the top-k spatial competition rule, as it takes into account both bottom-up feature inputs and top-down expectation inputs. The future studies will include general positions, variable sizes, and with-in-class object variations. More complete analysis with the model, in terms of memory efficiency and computational efficiency, will also be carried out.

References

[1] G. Backer, B. Mertsching, and M. Bollmann. Data- and model-driven gaze control for an active-vision system. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(12):1415–1429, December 2001.

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

[3] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1):107–155, 1962.

[4] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.

[5] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.

[6] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[7] M. Mozer and M. Sitton. *Attention*. UCL Press, London, 1996.

[8] B.A. Olshausen, C. Anderson, and D. 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.

[9] R. P. N. Rao and D. H Ballard. Probabilistic models of attention based on iconic representations and predictive coding. In L. Itti, G. Rees, and J. Tsotsos, editors, *Neurobiology of Attention*. Academic Press, 2004.

[10] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, and C. Zetzsche. Scene analysis with saccadic eye movemens: Top-down and bottom-up modeling. *Journal of Electronic Imaging*, 10(1):152–160, 2001.

[11] L. W. Stark and Y. S Choi. In W. H. Zangemeister, H. S. Stiehl, and C. Freska, editors, *Visual Attention and Cognition*, pages 3–96. Elsevier Science, Amsterdam, 1996.

[12] 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.

[13] 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.