Deep-BCN: Deep networks meet biased competition to create a brain-inspired model of attention control

Hossein Adeli
Stony Brook University
hossein.adelijelodar@stonybrook.edu

Gregory Zelinsky
Stony Brook University
gregory.zelinsky@stonybrook.edu

Abstract

The mechanism of attention control is best described by biased-competition theory (BCT), which suggests that a top-down goal state biases a competition among object representations for the selective routing of a visual input for classification. Our work advances this theory by making it computationally explicit as a deep neural network (DNN) model, thereby enabling predictions of goal-directed attention control using real-world stimuli. This model, which we call Deep-BCN, is built on top of an 8-layer DNN pre-trained for object classification, but has layers mapped to early visual (V1, V2/V3, V4), ventral (PIT, AIT), and frontal (PFC) brain areas that have their functional connectivity informed by BCT. Deep-BCN also has a superior colliculus and a frontal-eye field, and can therefore make eye movements. We compared Deep-BCN's eye movements to those made from 15 people performing a categorical search for one of 25 target object categories, and found that it predicted both the number of fixations during search and the saccade-distance travelled before search termination. With Deep-BCN a DNN implementation of BCT now exists, which can be used to predict the neural and behavioral responses of an attention control mechanism as it mediates a goal-directed behavior—in our study the eye movements made in search of a target goal.

1. Introduction

Visual object detection is important both for automated systems and biological systems. For computers, object detection underlies an untold number of applications, ranging from smart homes to self-driving cars. For biological systems, object detection is a core cognitive necessity that undoubtedly shapes a species' success, informing an animal when it is appropriate to fight or flee or forage. The importance of successful object detection has spawned separate but equally vast literatures. In computer vision, decades of vigorous research has culminated in international challenges aimed at improving automated object detection performance. A major advance of this literature is the development of increasingly precise and robust object detectors that can be passed over the pixels in an image or video. In biological vision, a half century of research by vision scientists, cognitive scientists, and neuroscientists produced massive literatures rich with data and theories. A major advance of this effort has been the development of a framework that is widely accepted among these researchers; that object detection is mediated by an attention control mechanism that biases the selective routing of visual inputs [1]. The computer vision and biological vision approaches to the object detection problem, although vastly different in their constraints, have an underlying similarity. Each assumes that object detection is a process of selectively analyzing local regions in an input, and repeating this analysis at multiple input locations. In computer vision this assumption takes the form of a moving-window object detector that is applied to all pixels in an image or, in a recent effort to improve efficiency, to only those pixels likely corresponding to objects [2]. In biological vision this assumption is epitomized by the metaphor of an attention "spotlight" that shifts from location to location, detecting objects in its path. Here we exploit this commonality and attempt using it to bridge these two richly evolved perspectives on the same object detection problem. We take this first step by melding an accepted behavioral framework for the primate attention control system with an artificial deep network architecture, and using this brain-inspired deep network to predict goal-directed behavior in humans.

1.1. Attention in Computers

The DNN literature is increasingly appealing to the concept of attention [3-6]. These networks learn how to weight the integration of visual inputs with trained internal representations in order to generate sequences of outputs. This sequential dynamic weighting mechanism has been shown to improve model performance across applications ranging from caption generation [5] to question answering

[7] and translation [8]. It has also been shown to improve multi-digit number recognition [4] by having a model attend to, and recognize, individual digits in a multi-digit string, thereby reducing the number of prediction classes to 10 (all the digits). Attention control has even been suggested as possibly replacing convolution and recurrent operations commonly used in current networks [8]. But the conceptualization of attention in these models maps only very loosely onto the detailed blueprint for how attention is implemented in the brain. Their architectures are also narrowly designed to perform specific tasks, making them lack what might be considered a general underlying attention mechanism. But if even a rough adoption of attention can prove useful across so many computer vision applications, what might models be able to do that adopt a more brain-inspired design?

1.2. Attention in the Brain

The core function of visual attention is to prioritize and select a subset of visual input for further processing, and the currently best theory for how this happens in the brain is Biased Competition (BC) ([9-15]. BC proposes that bottom-up visual information is weighted at various levels in a processing hierarchy by top-down modulations, with the goal of this biasing being to have task-relevant information win a neural competition for object recognition and the control of goal-directed behavior.

Attentional biases can be both spatial [16, 17] and for the features of objects [18, 19]. The consequence of attention biasing a location in space is to cause information about an object appearing at that location to be selectively routed to higher visual areas. Mechanistically, this is accomplished by shrinking the effective receptive fields (RFs) of neurons to the attended object, as if that object was presented in isolation [20]. The behavioral expression of this *selective routing* of information is a facilitated interaction with an object, presumably because it has been recognized, and the consequent reduced ability to interact with unattended objects. Feature and object biases work similarly, except that the routed information relates to an object's features rather than its spatial location. These biases therefore modulate neural responses in parallel, and are not limited to a single area of focus [21]. Feature biases are known to originate from pre-frontal cortex (PFC), and exert their influence via modulation of frontal eye field (FEF) activity [22], which in turn biases activity in mid-level visual areas. Object biases similarly originate in PFC, but exert their influence by feeding back to modulate activity in higher visual areas [23]. In big picture, BC proposes that a feature bias corresponding to an object goal introduces a spatial bias that selectively routes visual inputs through the visual pathways for the purpose of mediating classification of and interaction with the object.

1.3. Models of Biased Competition

Many models of visual attention share the conceptual framework of BC. Some of these models are largely mathematical, such as the Neural Theory of Visual Attention (NTVA) [24-26]. NTVA is a large-scale model of the brain areas comprising the attention network, and captures the top-down feature and spatial-biasing processes and the competition mechanisms theorized by BC. However, NTVA inputs probability distributions associated with features of object categories and does not extract these features from pixels or learn categories from image exemplars. It is therefore not a computational model according to at least one widely-accepted definition [27]. Other models *are* computational, and borrow heavily from methods developed in computer vision. Perhaps the best computational model of general attention is Selective Tuning (ST) [28, 29]. ST is another large-scale model employing multiple mechanisms, but its focus is on the neural modulation underlying selectivity and not the mapping between specific brain areas in the attention network and the prediction of explicit behavior in response to complex stimuli. Another class of attention models [30-32] sticks closer to the neurophysiology of brain structures and attempts to capture the network dynamics missing from ST, but these models also tend to be smaller in scale, focusing on interactions between only a small subset of the brain areas in the attention network.

1.4 Ventral Visual Processing and Convolutional Neural Networks

The detection and interaction with visual objects in naturalistic tasks requires the formation and use of rich object category representations that activate in response to a visual input. This is true for both biological and computer systems. In primates this is accomplished by processing along the ventral visual pathway of cortical brain structures [33, 34]. The ventral, or "what", pathway starts from the primary visual area, V1, in the occipital cortex and extends temporally to V2/V3, V4, Posterior inferotemporal (PIT), and ultimately Anterior inferotemporal (AIT) area in the inferotemporal cortex (IT). Processing along this ventral pathway endows primates with an ability to recognize objects and scenes [33, 35-37]. Moreover, this processing is hierarchical; neurons in early areas code for low-level visual features (e.g., orientation), and those in higher areas selectively code complex visual patterns and categories (e.g., faces).

In computer vision, the best representations for mediating object detection are learned using artificial deep neural networks (DNNs). DNNs have been remarkably successful in their ability to recognize objects and scenes

[38, 39], easily surpassing previous methods, and this has resulted in them dominating the computer vision and machine learning literatures [39]. DNNs are a class of models, with perhaps the most popular being Convolutional Neural Networks (CNNs) [40]. CNNs were inspired by the hierarchical architecture of the mammalian cortex [40, 41], mainly the ventral pathway of visually-responsive brain areas. A typical CNN processes an image, first through several convolutional layers, then through fewer fully-connected layers, and finally through a classification layer linked to some decision or action. The nodes of the convolutional layers consist of filters, with each taking as input the outputs of a subset of nodes at the lower layer. The hierarchical convolution of filters with an image mimics the parallel extraction of information over visual space performed by hypercolumns in the early visual areas [42]. The power of these deep networks lies in the feature representations learned across their layers, each richer than the one below, another property paralleling the organization of structures along the ventral pathway.

Given their brain inspiration and stunning success, CNNs are beautifully suited to model ventral pathway processing in the primate brain. This exciting prospect has not gone unnoticed. Recent work has compared the selectivity and accuracy of representations built across a DNN's layers to those of brain areas in the ventral pathway [43-49], where it was shown that the final layer of a CNN trained on object classification can predict neural responses in IT cortex [44, 45] and that both classification accuracy and filter selectivity in the intermediate network layers capture neural selectivity along the intermediate ventral pathway [43]. To date, however, there has been no attempt to design a DNN to reflect the brain connectivity and interactions between structures in the broader attention network.

In this paper we take brain inspiration to the next level by integrating the principles of biased competition into the architecture of a DNN. This model, which we call Deep-BCN (Deep Biased Competition Network) is the first DNN to use attention-inspired modulations of network activation to predict the goal-directed behavior of humans performing a task. The task that we chose is visual search, as this is the simplest and clearest example of a goal-directed behavior (people looking for an instructed target object goal). The specific behaviors that we will predict are the eye movements made while people search, chosen because eye movements are observable behavioral expressions of individual shifts of attention [50]. Eye movements are also known to be guided to search targets under certain conditions [51, 52], with this *target guidance* being a measurable expression of the top-down feature biasing posited by BC theory. Our chosen combination of task and behavior are therefore well suited for evaluating Deep-BCN.

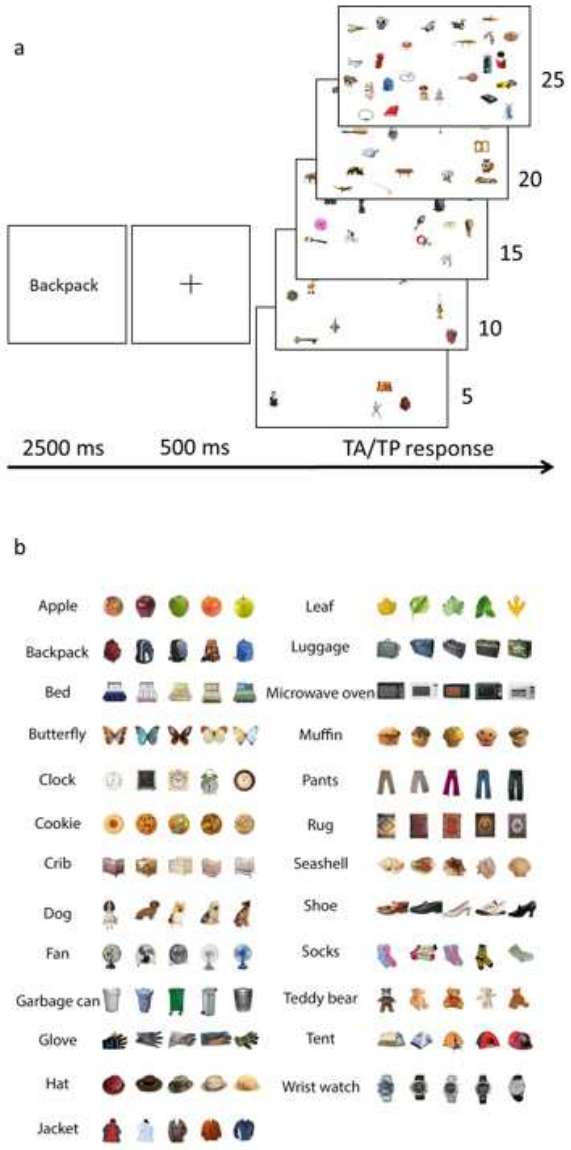

Figure 1: The categorical search experimental paradigm used by [53]. **(b)** The 125 category exemplars appearing as targets in the search images, and the 25 names designating the target category. Each of the 5 search set size conditions used a different target from the 5 target exemplars per category.

2. Behavioral Methods

The behavioral data used in this study were from [53], and that paper should be consulted for a detailed description of the methods. Briefly, 15 Stony Brook University undergraduates searched through arrays of common objects (Hemera Technologies) for each of 25 target categories while their eye movements were recorded (EyeLink 1000, SR Research; tower-mount configuration).

For each of these 25 categories, 5 highly typical exemplars were selected as targets to appear in the search displays [54]. Figure 1 shows these 25 target categories and the specific exemplars used as targets, as well as the experimental procedure. Search displays were constructed by randomly placing objects on a white background. These object arrays were used instead of realistic scenes so as to be able to manipulate the number of objects in the search image (*set size*) and to avoid the biasing of attention (gaze) by scene context, a topic beyond the scope of this initial study. Specifically, search images consisted of 5, 10, 15, 20, or 25 objects (5 levels of set size), with the constraints that objects could not overlap and no object could appear within 2° of the display center, a location corresponding to starting fixation. Each search display subtended a 47° horizontal by 28° vertical visual angle, based on a 1680×1049 pixel image viewed at 57 cm (a distance fixed by chinrest). Half of these displays were *target-present*, meaning that one of the depicted objects was an exemplar from the designated target category, and the other objects were *distractors*. Distractors were 3,700 objects from various categories, selected to comprise a disjoint set from the 25 target object categories. The other half of the displays were *target-absent*, consisting entirely of distractors. Targets and distractors subtended an average visual angle of 3.5° by 3.5°, and no object appeared twice throughout the experiment.

Each trial began by cuing a participant with the name of the target category (2,500 ms), followed by a central fixation cross (500 ms) and finally by presentation of a search display. Participants indicated their target-present or target-absent judgment by pressing either the right or left triggers of a game pad, respectively. Accuracy feedback was not provided. There were 10 practice trials and 250 experimental trials, each corresponding to a unique search image, and the experimental trials were evenly divided into the 5 set sizes and target-present and target-absent conditions, leaving 25 trials per cell of the design.

3. Model Methods

Figure 2 shows the anatomy of Deep-BCN. Deep-BCN is a DNN whose structure and connectivity are informed by the known neurophysiology and anatomy of the visual attention network [23, 55]. Core to Deep-BCN is its layers extending ventrally (the lower boxes in Fig. 2) corresponding to the ventral pathway of brain areas (V1, V2/V3, V4, PIT, and AIT) known to be important for visual object recognition in primates. We model these ventral brain areas using the 8-layer CNN known as AlexNet [56]. AlexNet has 5 convolutional and 3 fully-connected layers and is trained on recognizing 1000 object categories from the ImageNet dataset [57]. A pre-trained AlexNet was fine-tuned to recognize the 25 object categories from the behavioral experiment. This training dataset consisted of 1000 images of objects from each category and was gathered from ImageNet and other online sources. This fine-tuned AlexNet achieved an 88% level of recognition accuracy for the target objects used in the search displays when these objects were presented in isolation.

3.1. Early Bottom-up Visual Processing

The pipeline of Deep-BCN is as follows, loosely translated into hypothesized processing by the brain areas in the attention network. An image of a search display from the behavioral experiment is input to the fine-tuned network and processed by 5 feed-forward convolutional layers, corresponding to processing by early visual areas V1, V2/V3, and V4. Neurons in these areas are known to be selective to low-level visual features (e.g., orientation, color, intensity) and have antagonistic receptive field (RF) organizations similar to the responses from early layer convolutional filters trained for object classification [58, 59]. The sizes of neuron RFs also increase with movement along the early ventral visual stream, with neurons in V1 having the smallest RFs and neurons in V4 having larger RFs. This, too, parallels the architecture of a CNN, where filters at lower convolutional layers have a smaller RF size than those at higher layers [59]. And perhaps most importantly for our study, neuron RFs organize themselves into what is known as a retinotopic map of visual space [60]. Convolutional processing maintains a comparable retinotopic spatial organization of the input image, and in Deep-BCN a coarse retinotopy is preserved up to the fifth convolutional "V4" layer. This V4 retinotopy is crucial to our model, as it is here that representations are spatially biased for the purpose of improving object detection.

Another important property of primate retinotopic organization is that inputs are progressively blurred with increasing distance (or *eccentricity*) from a central region of visual space known as the fovea. Inputs arriving at our central vision are in high resolution, whereas those arriving in our peripheral vision are in lower resolution. In Deep-BCN we capture this eccentricity-dependent coding of resolution using a grid of 17×11 networks (Fig 3a), each of which "sees" a differently-sized and differently-localized patch of the larger input image. Specifically, the input to the highest resolution central network was a 100×100 pixel crop of the input image surrounding Deep-BCN's current fixation location, approximating a fovea that can be directed to only a single object. The sizes of image crops increased with distance away from the center, up to a maximum size of 450×450 pixels for networks farthest in the periphery (i.e., those covering the corners of the input image in this example).

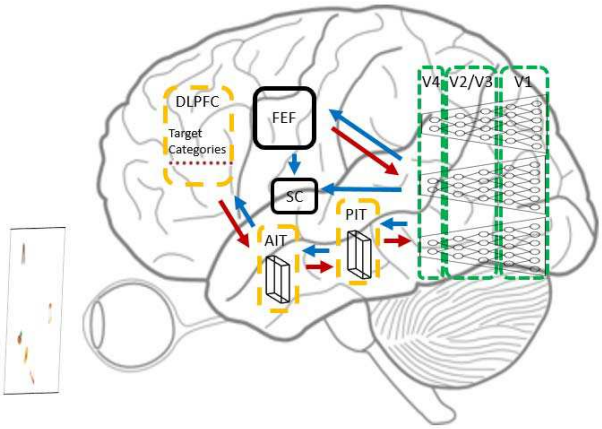

Figure 2: Anatomy of Deep-BCN. Dashed boxes indicate layers of a CNN, with green coding convolutional layers and yellow coding fully-connected layers. Nodes in the last layer, labeled DLPFC, correspond to the 25 target categories learned by the network. The solid box labeled FEF selects a winner in the biased V4 activation. The solid box labeled SC indicates a non-DNN model from [53]. Blue arrows indicate feedforward connections and red arrows indicate feedback connections. Note that several other connections between these brain areas are known to exist but are not shown so as to clearly specify only those connections implemented in Deep-BCN. DLPFC = dorsolateral pre-frontal cortex, FEF = frontal eye field, SC = superior colliculus, and AIT and PIT = anterior and posterior inferotemporal cortex, respectively.

Figure 3b shows samples of how the sizes of Deep-BCN's RFs change with distance from its fovea. Note that the size of the foveal RF limits the cropped region to only the fixated object (the teddy bear), while the larger peripheral RFs result in crops often depicting two or more objects. The consequence of this is that the features extracted from peripheral regions are often aggregated over multiple objects, leading to the poorer classification of any given one, while features extracted foveally will be limited to a single object and therefore will likely yield a higher classification confidence. Each of the 187 (11×17) networks in the grid is input a RF-cropped image patch and outputs a 13×13×256 tensor, which we will treat as 13×13 pixel activation maps for each of 256 features. These eccentricity-dependent activation maps are then averaged to create a single activation map for each feature dimension. Taking the average of these 256 feature activation maps, then multiplying the foreground pixel values to isolate activity to the objects, gives us a V4 (layer 5) activation map corresponding to Deep-BCN's estimate of purely bottom-up spatial priority. According to BCT, it is the weighting of features in this combined priority map that can be biased by top-down control.

Areas PIT and AIT are inferotemporal structures known to mediate object recognition in primates, and these correspond to layers 6 and 7 in Deep-BCN. The projection

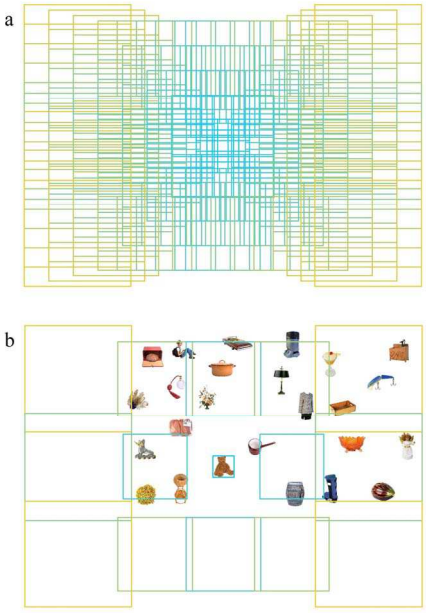

Figure 3: An eccentricity-dependent network of networks. **(a)** The 187 regions input to the 11× 17 grid of networks used to tile the larger region of space corresponding to the full search image. **(b)** Subset of networks illustrating the increase in "receptive field" size that occurs with increasing eccentricity relative to the high-resolution central network.

from V4 (layer 5) to PIT (layer 6) is modeled as fully connected, and this is also the case for the projection from PIT to AIT (layers 6 to 7). This implementation decision is informed by the fact that neurons in these IT structures have RFs covering much of the visual field, the end product of retinotopy largely disappearing in these late ventral visual areas. The consequence of this loss of retinotopy is that the nodes in Deep-BCN's layers must compete for exclusive access to the RFs of nodes at the next higher level. This bottom-up competition selects a winning RF, and it is the location of this RF in space that determines the selective routing of visual inputs through to IT and, ultimately, the classification of that "attended" input as a category of object. This classification is performed by a classification layer (layer 8), presumably existing in the Dorsolateral PFC (DLPFC), which maps the outputs of the IT nodes to the 25 category nodes that have been learned by the network. Each category is a node in this layer, and it is through the connections between these DLPFC nodes and the rich representations formed in IT from earlier ventral processing that patterns in image inputs are classified by Deep-BCN.

3.2. Top-down Attention Control and Selective Routing

A top-down mechanism of attention control is essential for the production of any goal-directed behavior. Behavior

might exist in its absence, but it would be controlled purely by the input and not by the goal of performing a task. In the task of visual search, the DLPFC has been implicated in the creation of "templates" of goal states in visual working memory, for the purpose of biasing the bottom-up competition for neural resources and the selective routing of visual inputs for classification [61]. In Deep-BCN, the DLPFC nodes serve a dual function: they constitute the set of learned categories against which inputs will be evaluated, and they are the source of biases for specific target-object goals. Our assumption is that the word cue used in the categorical search paradigm would activate the corresponding DLPFC node in Deep-BCN, and this in turn would feed activation back down the ventral stream so as to modulate activity in the IT and V4 layers. For the current first approximation of a systems-level model of the attention control mechanism we implement only a subset of the feedback projections known to exist throughout the ventral pathway [55]. Specifically, we implement projections from DLPFC to IT, and from IT to V4. We did this because areas V4 and higher have been shown to be most instrumental for target selection within a biased-competition framework [62-64]. According to Deep-BCN, attention control is exerted by feedback from the fully-connected layers biasing the bottom-up V4 activations of the rich feature representations formed in the last convolutional layer.

The top-down attention control signal is modeled using the same connections as in the feedforward projections, but using the gradient signal. In the supervised training of a DNN, the gradient signal is used to modify the weights between layers so as to move the classification response in the desired direction. Under Deep-BCN, the DLPFC node corresponding to the target goal exerts a gradient signal that changes the filter weightings so as to favor the prediction of the target category (implemented using Grad-CAM [65]). Stated differently, the gradient signal originating from a classification node modulates the gain of the filters that respond best to features of the target goal. We combine this top-down bias with bottom-up processing by simply multiplying the gradient feedback weighting and the feedforward (bottom-up) activation, thereby biasing the competition between nodes at these layers in favor of the target's features.

Such gain modulation is widely studied under the topic of *feature-based attention* [21, 62, 63, 66], and in Deep-BCN it is the reentrant interaction between FEF and V4 [67] that resolves the competition for RFs. Note that the connections between FEF and V4 are not trainable, meaning that FEF is not a fully-integrated part of the ventral CNN. Rather, it is a relatively independent module that selects a winning routing window from the biased V4 activation map. This feature biasing and competition resolution causes the selective routing of visual inputs at a region of space, and this leads to the constriction of IT RFs to that region and the possible detection of a pattern in those select inputs as a member of the cued object category. Classification is therefore attempted on the output of the most active network from the early visual processing grid.

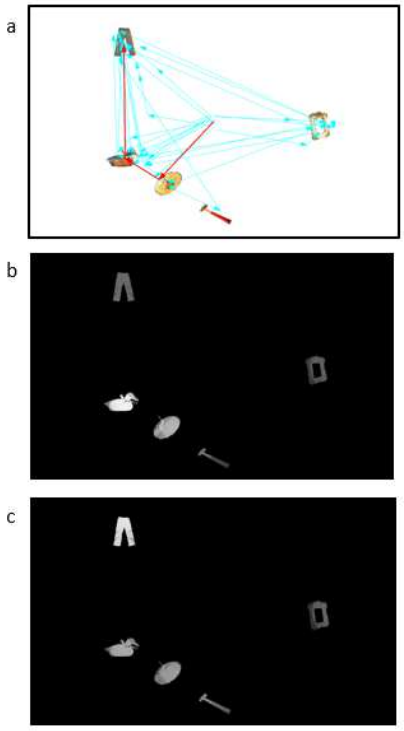

Figure 4: Representative activation patterns in Deep-BCN underlying its shifts of gaze to a target goal. **(a)** Input is an image of an object array. Scanpaths from subjects (cyan) and the model (red) for a categorical "pants" search task. The subjects and model started each trial fixated at the center. **(b)** The bottom-up retinotopic activation in V4 in response to the input image. **(c)** The biased V4 activation map after multiplying the bottom-up activation and the top-down activation feeding back from the "pants" frontal node. Note that modulation of the target-like features leads to increased activation at the locations of the target-like objects, seen most clearly for the exemplar of the pants target.

Finally, Deep-BCN has a fovea that it moves from location to location in the search image while attempting to detect the target object category, much like people do. We believe that the importance of this eye movement behavior in object detection success is often understated. This neglect is certainly apparent in BC studies of attention control during visual search [21]. But while not part of the core BC framework, eye movements are behavioral expressions of the same BC operations [21]. In the context of our behavioral search task, at any given fixation a new eye movement might be programmed to any of the objects

appearing in the search image. The locations of these objects therefore compete for selection, with the winning location determining the next saccade vector. This competition, often conceptualized and studied as priority maps [68], likely involves a whole other network of frontal-parietal structures [69, 70] that we grossly simplify here to just the FEF. Under Deep-BCN, this spatial bias is also the product of the DLPFC→IT→V4 feedback bias projecting from V4 to FEF (the feedforward connection between the two in Figure 2). One could think of this bias as indicating a target/non-target classification confidence score attached to the locations of the networks (in the grid of networks) covering the region of space subtended by the search image, and FEF activity corresponding to the selection of a winning network in the biased V4 activation.

The FEF then communicates this bias to the superior colliculus (SC), a mid-brain structure implicated in the production of oculomotor behavior, where it combines with the biased retinotopic activity projecting from V4 (Figure 2) to create in the SC a priority map for the express purpose of controlling saccades (often referred to as *overt* shifts of attention). The output of this behavior-scale BC model is a sequence of eye movements, each one aligning the center of the eccentricity-dependent grid of networks with the location of the to-be-classified pattern. This behavior therefore serves a similar function as selective routing under a BC framework; rather than a purely neural gating of early visual inputs through to classification, now inputs are gated in the sense that high-resolution foveal processing is being brought to bear on inputs at a select region of visual space. But regardless of whether this attention-controlled gating is covert or overt, both use the same cost function—the maximization of classification success. At the level of the SC, Deep-BCN uses MASC (Model of Attention in the Superior Colliculus) [53] to generate saccades. MASC inputs a priority map, here the equally-weighted integration of the descending V4 and FEF activity, and after processing informed by known collicular architecture and neurophysiology, outputs a coordinate coding the next fixation location (the original text should be consulted for details). Target-present searches were terminated when the model classified a fixated object as an exemplar of the target category with an 80% level of confidence, which was a parameter of our model. Search was terminated with a target-absent judgement if activation in the V4 layer fell below a set threshold, which was tuned to get the best fit to the behavioral data in the target-absent condition.

4. Results

In an initial qualitative evaluation of Deep-BCN we wanted to determine two things: does its attention control mechanism select reasonable "human-like" image locations to route its visual inputs (as opposed to patently artificial movements of spatial attention), and is the bias originating from its learned target categories sufficiently large to affect activity at the V4 layer after its back-projection from the DLPFC nodes?

The answer to the first question is clear from Figure 4a, which shows representative eye movement behavior from Deep-BCN and participants superimposed over the object array being searched. The target category of "pants" was designated to subjects using a word cue appearing immediately before the search display, and was designated to Deep-BCN by activation of its DLPFC node corresponding to its "pants" category. Far from being artificial, Deep-BCN shifted its attention to reasonable image locations in the search for its goal. Indeed, had its behavior not been coded in a different color it would likely be indistinguishable from the subject behavior shown in cyan. Figure 5 gives a clearer sense of how a representative *scanpath* of attention shifts from the model compared to six randomly selected scanpaths of individual participants searching the same image, this time for a "glove" target. Once again, Deep-BCN's scanpath of eye movements agreed well with those of subjects, both in terms of their number and in their trajectory. Such qualitative similarities are an often overlooked but important dimension to consider in a model evaluation of this type; it may be the case that a model generates reasonably good agreement to human behavior in terms of various quantitative metrics, but distinctly poor agreement when the scanpaths of the two are visualized and compared. Deep-BCN passed this initial test.

As for the question of whether Deep-BCN's behavior reflects a V4 bias, this answer is given in the comparison between Figure 4b and Figure 4c for a "pants" search target. Specifically, Figure 4b shows bottom-up retinotopic activity from V4 before it has been biased from the DLPFC "pants" node. Note that activity corresponding to

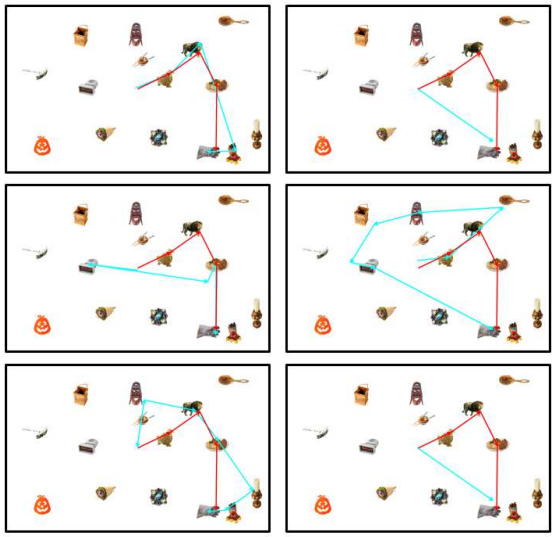

Figure 5: Representative scanpaths from Deep-BCN (red) and subjects (cyan) for a categorical "glove" search task.

the pants object is less than the activity elicited by some of the non-targets. Figure 4c shows this same bottom-up V4 activity combined with the top-down DLPFC bias. Now the pants are more active. This difference is due to the fact that feature maps are weighted equally in the unbiased case (Fig. 4b) but are weighted using the gradient signal from the "pants" node in the biased case (Fig. 4c). This demonstrates that Deep-BCN is able to capture the core component of BCT, a biasing of the visual input for the purpose of achieving a behavioral goal.

Turning to a more quantitative evaluation, analysis of the behavioral data revealed patterns that are now classic in the visual search literature. As shown in Figure 6, search became more difficult with increasing set size, evidenced by an increase in the number of gaze fixations and the total distance traveled by gaze during search (summed saccade-vector-length distance). The cyan functions show the behavioral means for the number of fixations and distance-travelled metrics, grouped by set size and images in which an exemplar of the target category was present (6a) or absent (6b). For both metrics, behavior was analyzed up to the button press terminating the target-present or target-absent trial. These patterns are the behavioral ground truth for attention control against which Deep-BCN was evaluated. Because Deep-BCN also makes eye movements in its search for a target, its behavior in response to the same images can be grouped into the same conditions and evaluated using the same metrics, thereby enabling a direct comparison to participant behavior. As shown, Deep-BCN not only predicted the effect of set size on both number of fixations and distance travelled, it also predicted the effect of target presence/absence and its interaction with set size (search was more affected by increasing set size in the target-absent images compared to the target-present). Although the two behavioral metrics are highly correlated, both capturing the same breakdown in the ability of the attention control mechanism to selectively route inputs from image locations corresponding to the target category goals, the fact that Deep-BCN captured so well the effects of set size and target presence/absence on this breakdown is nevertheless impressive, even more so because it did this by virtue of its inclusion of basic biased-competition components of the primate attention control mechanism.

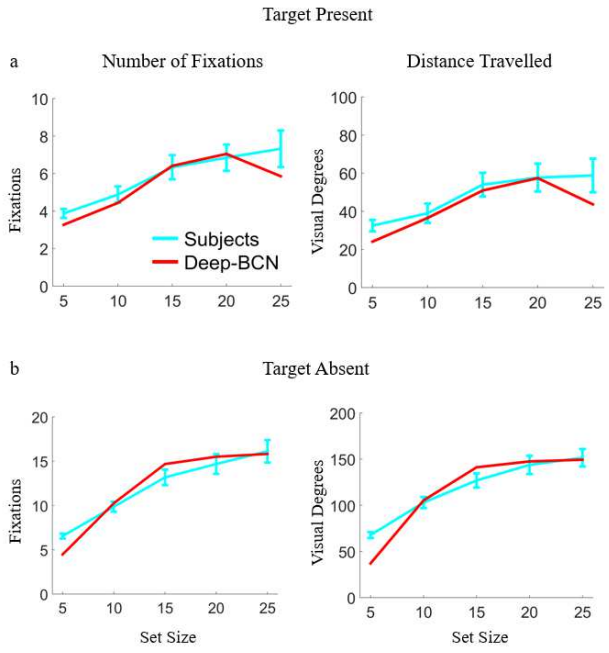

Figure 6: Comparison of Deep-BCN and participants' mean eye movement behavior in the categorical search dataset. (a) Plots showing the number of fixations and distance traveled to the target (summed saccade distance) for Deep-BCN (red) and all participants (cyan) as a function of set size for the target-present images. (b) Similar plots for the target-absent images. For both metrics, data is shown up to participants ending a trial with a target present/absent button press response or the model meeting its target present/absent termination criteria. Error bars plot the SEM.

5. Discussion

The computer vision literature is recognizing the importance of including a mechanism of attention control into its methods. Here we took the core tenets of biased-competition theory, that attention control is the top-down biasing of visual inputs for the purpose of achieving some behavioral goal, and built these into a deep neural network model that we call Deep-BCN. Deep-BCN contributes to the computer vision literature in making far more explicit the concept of attention control. To the extent there is value in the inclusion of an attention mechanism in computer vision methods, this value may increase as the implemented mechanism of attention more closely matches the mechanism implemented in the brain. One direction for future work will be to conduct a model comparison to see whether deep networks engineered after the primate brain outperform comparable state-of-the-art object detectors that are less brain-inspired, thus informing the value of neuroengineering computational methods. Deep-BCN contributes to the behavioral vision literature in being the first deep network model of a goal-directed behavior, specifically the eye movements leading up to the target detection decision in a search task. We believe that Deep-BCN's success in predicting this goal-directed human behavior stems from its BCT-inspired design. But Deep-BCN is just a gross first-approximation attempt to design into a DNN a mechanism of attention control, and another important direction for future work will be to see whether more brain-inspired designs will lead to even better predictions of goal-directed human behavior.

References

[1] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience,* vol. 18, pp. 193-222, 1995.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91-99.

[3] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems*, 2014, pp. 2204-2212.

[4] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755,* 2014.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov*, et al.*, "Show, attend and tell: Neural image caption generation with visual attention," *arXiv preprint arXiv:1502.03044,* vol. 2, p. 5, 2015.

[6] Z. Wei, H. Adeli, G. Zelinsky, M. Hoai, and D. Samaras, "Learned Region Sparsity and Diversity Also Predict Visual Attention," presented at the NIPS, 2016.

[7] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani*, et al.*, "Ask me anything: Dynamic memory networks for natural language processing," in *International Conference on Machine Learning*, 2016, pp. 1378-1387.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez*, et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000-6010.

[9] J. Duncan, "Cooperating brain systems in selective perception and action," 1996.

[10] G. R. Mangun and S. A. Hillyard, "Modulations of sensory-evoked brain potentials indicate changes in perceptual processing during visual-spatial priming," *Journal of Experimental Psychology: Human perception and performance,* vol. 17, p. 1057, 1991.

[11] J. H. Maunsell and S. Treue, "Feature-based attention in visual cortex," *Trends in neurosciences,* vol. 29, pp. 317-322, 2006.

[12] D. M. Beck and S. Kastner, "Top-down and bottom-up mechanisms in biasing competition in the human brain," *Vision research,* vol. 49, pp. 1154-1165, 2009.

[13] M. Saenz, G. T. Buracas, and G. M. Boynton, "Global effects of feature-based attention in human visual cortex," *Nature neuroscience,* vol. 5, pp. 631-632, 2002.

[14] J. T. Serences and G. M. Boynton, "Feature-based attentional modulations in the absence of direct visual stimulation," *Neuron,* vol. 55, pp. 301-312, 2007.

[15] M. A. Schoenfeld, J.-M. Hopf, C. Merkel, H.-J. Heinze, and S. A. Hillyard, "Object-based attention involves the sequential activation of feature-specific cortical modules," *Nature neuroscience,* vol. 17, pp. 619-624, 2014.

[16] S. Kastner and L. G. Ungerleider, "Mechanisms of visual attention in the human cortex," *Annual review of neuroscience,* vol. 23, pp. 315-341, 2000.

[17] S. J. Luck, L. Chelazzi, S. A. Hillyard, and R. Desimone, "Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex," *Journal of neurophysiology,* vol. 77, pp. 24-42, 1997.

[18] S. Kastner, P. De Weerd, and L. G. Ungerleider, "Texture segregation in the human visual cortex: a functional MRI study," *Journal of Neurophysiology,* vol. 83, pp. 2453-2457, 2000.

[19] M. Stokes, R. Thompson, A. C. Nobre, and J. Duncan, "Shape-specific preparatory activity mediates attention to targets in human visual cortex," *Proceedings of the National Academy of Sciences,* vol. 106, pp. 19569-19574, 2009.

[20] L. Chelazzi, J. Duncan, E. K. Miller, and R. Desimone, "Responses of neurons in inferior temporal cortex during memory-guided visual search," *Journal of neurophysiology,* vol. 80, pp. 2918-2940, 1998.

[21] N. P. Bichot, A. F. Rossi, and R. Desimone, "Parallel and serial neural mechanisms for visual search in macaque area V4," *Science,* vol. 308, pp. 529-534, 2005.

[22] N. P. Bichot, M. T. Heard, E. M. DeGennaro, and R. Desimone, "A source for feature-based attention in the prefrontal cortex," *Neuron,* vol. 88, pp. 832-844, 2015.

[23] D. Baldauf and R. Desimone, "Neural mechanisms of object-based attention," *Science,* vol. 344, pp. 424-427, 2014.

[24] C. Bundesen, T. Habekost, and S. Kyllingsbæk, "A neural theory of visual attention: bridging cognition and neurophysiology," *Psychological review,* vol. 112, p. 291, 2005.

[25] C. Bundesen, "A theory of visual attention," *Psychological review,* vol. 97, p. 523, 1990.

[26] C. Bundesen, T. Habekost, and S. Kyllingsbæk, "A neural theory of visual attention and short-term memory (NTVA)," *Neuropsychologia,* vol. 49, pp. 1446-1457, 2011.

[27] J. Tsotsos and A. Rothenstein, "Computational models of visual attention," *Scholarpedia,* vol. 6, p. 6201, 2011.

[28] J. Tsotsos, "Analyzing vision at the comlexity level.—Behav," *Brain Sei,* vol. 13, pp. 423-469, 1990.

[29] J. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial intelligence,* vol. 78, pp. 507-545, 1995.

[30] F. H. Hamker, "The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement," *Cerebral cortex,* vol. 15, pp. 431-447, 2005.

[31] F. H. Hamker, "A dynamic model of how feature cues guide spatial attention," *Vision research,* vol. 44, pp. 501-521, 2004.

[32] G. Deco and J. Zihl, "Top-down selective visual attention: A neurodynamical approach," *Visual Cognition,* vol. 8, pp. 118-139, 2001.

[33] L. G. Ungerleider and J. V. Haxby, "'What'and 'where'in the human brain," *Current opinion in neurobiology,* vol. 4, pp. 157-165, 1994.

[34] L. G. Ungerleider and L. Pessoa, "What and where pathways," *Scholarpedia,* vol. 3, p. 5342, 2008.

[35] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral cortex,* vol. 1, pp. 1-47, 1991.

[36] L. G. Ungerleider, "Two cortical visual systems," *Analysis of visual behavior,* pp. 549-586, 1982.

[37] D. J. Kravitz, K. S. Saleem, C. I. Baker, L. G. Ungerleider, and M. Mishkin, "The ventral visual pathway: an expanded neural framework for the processing of object quality," *Trends in cognitive sciences,* vol. 17, pp. 26-49, 2013.

[38] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487-495.

[39] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, pp. 436-444, 2015.

[40] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard*, et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural computation,* vol. 1, pp. 541-551, 1989.

[41] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics,* vol. 36, pp. 193-202, 1980.

[42] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology,* vol. 160, pp. 106-154, 1962.

[43] U. Güçlü and M. A. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *The Journal of Neuroscience,* vol. 35, pp. 10005-10014, 2015.

[44] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon*, et al.*, "Deep neural networks rival the representation of primate IT cortex for core visual object recognition," *PLoS Comput Biol,* vol. 10, p. e1003963, 2014.

[45] D. Yamins, H. Hong, C. Cadieu, and J. J. DiCarlo, "Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream," in *Advances in neural information processing systems*, 2013, pp. 3093-3101.

[46] D. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences,* vol. 111, pp. 8619-8624, 2014.

[47] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," *Neuron,* vol. 73, pp. 415-434, 2012.

[48] S.-M. Khaligh-Razavi and N. Kriegeskorte, "Deep supervised, but not unsupervised, models may explain IT cortical representation," *PLoS Comput Biol,* vol. 10, p. e1003915, 2014.

[49] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, "Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition," *arXiv preprint arXiv:1601.02970,* 2016.

[50] M. Shepherd, J. M. Findlay, and R. J. Hockey, "The relationship between eye movements and spatial attention," *The Quarterly Journal of Experimental Psychology Section A,* vol. 38, pp. 475-491, 1986.

[51] G. J. Zelinsky, H. Adeli, Y. Peng, and D. Samaras, "Modelling eye movements in a categorical search task," *Phil. Trans. R. Soc. B,* vol. 368, p. 20130058, 2013.

[52] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Image processing, 2003. icip 2003. proceedings. 2003 international conference on*, 2003, pp. I-253.

[53] H. Adeli, F. Vitu, and G. J. Zelinsky, "A model of the superior colliculus predicts fixation locations during scene viewing and visual search," *Journal of Neuroscience,* vol. 37, pp. 1453-1467, 2017.

[54] T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva, "Conceptual distinctiveness supports detailed visual long-term memory for real-world objects," *Journal of Experimental Psychology: General,* vol. 139, p. 558, 2010.

[55] C. D. Gilbert and W. Li, "Top-down influences on visual processing," *Nature Reviews Neuroscience,* vol. 14, pp. 350-363, 2013.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 248-255.

[58] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014, pp. 818-833.

[59] D. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature neuroscience,* vol. 19, pp. 356-365, 2016.

[60] S. A. Engel, G. H. Glover, and B. A. Wandell, "Retinotopic organization in human visual cortex and the spatial precision of functional MRI," *Cerebral cortex (New York, NY: 1991),* vol. 7, pp. 181-192, 1997.

[61] C. N. Olivers, J. Peters, R. Houtkamp, and P. R. Roelfsema, "Different states in visual working memory: When it guides attention and when it does not," *Trends in cognitive sciences,* vol. 15, pp. 327-334, 2011.

[62] J. H. Reynolds, T. Pasternak, and R. Desimone, "Attention increases sensitivity of V4 neurons," *Neuron,* vol. 26, pp. 703-714, 2000.

[63] J. H. Reynolds and R. Desimone, "Interacting roles of attention and visual salience in V4," *Neuron,* vol. 37, pp. 853-863, 2003.

[64] L. Chelazzi, E. K. Miller, J. Duncan, and R. Desimone, "A neural basis for visual search in inferior temporal cortex," *Nature,* vol. 363, p. 27, 1993.

[65] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *See* https://arxiv. *org/abs/1610.02391 v3,* 2016.

[66] H. Zhou and R. Desimone, "Feature-based attention in the frontal eye field and area V4 during visual search," *Neuron,* vol. 70, pp. 1205-1217, 2011.

[67] G. G. Gregoriou, S. J. Gotts, H. Zhou, and R. Desimone, "High-frequency, long-range coupling between prefrontal and visual cortex during attention," *science,* vol. 324, pp. 1207-1210, 2009.

[68] L. Itti and A. Borji, "Computational models: Bottom-up and top-down aspects," *arXiv preprint arXiv:1510.07748,* 2015.

[69] D. Schluppeck, C. E. Curtis, P. W. Glimcher, and D. J. Heeger, "Sustained activity in topographic areas of human posterior parietal cortex during memory-guided saccades," *Journal of Neuroscience,* vol. 26, pp. 5098-5108, 2006.

[70] S. Kastner, K. DeSimone, C. S. Konen, S. M. Szczepanski, K. S. Weiner, and K. A. Schneider, "Topographic maps in human frontal cortex revealed in memory-guided saccade and spatial working-memory tasks," *Journal of neurophysiology,* vol. 97, pp. 3494-3507, 2007.