Visual Search Asymmetry: Deep Nets and Humans Share Similar Inherent Biases

Anonymous Author(s)
Affiliation
Address
email

Abstract

Visual search is a ubiquitous and often challenging daily task, exemplified by looking for the car keys at home or a friend in a crowd. An intriguing property of some classical search tasks is an asymmetry such that finding target A among distractors B can be easier than finding B among A. En route towards a quantitative framework that accounts for the mechanisms responsible for asymmetry in visual search, we propose a computational model that takes a target and a search image as inputs and produces a sequence of eye movements. The model integrates eccentricity-dependent visual recognition with target-dependent top-down cues. We tested and compared the model against human behavior in six paradigmatic search tasks that produce asymmetries in humans. Without any prior exposure to the stimuli or any task-specific training, the model provides a plausible mechanistic basis underlying asymmetric search. Inspired by the similar behavior of models and humans, we investigated the origins of asymmetric search in the model. We hypothesized that the polarity of search asymmetry arises from experiences in the natural environment. We tested this hypothesis by training the model on an augmented version of ImageNet where the original biases of the natural images are either removed or reversed. The results showed that the polarity of search asymmetry disappeared or was altered depending on the training protocol. This study highlights how classical perceptual properties can emerge in neural network models, without the need for task-specific training, but rather as a natural outcome of the statistical properties of the developmental diet fed to the model.

1 Introduction

Humans and other primates continuously move their eyes in search of objects, food, or friends. Extensive psychophysical studies have documented how these visual search behaviors depend on the complex interplay between the nature of the target objects, the search images, and the subjects' memory and attention [26, 13, 18, 22, 2, 8, 14]. There has also been significant progress in describing the neurophysiological steps involved in processing visual information [16, 20, 11] and the neuronal circuits that orchestrate attention and eye movements [5, 19, 15, 3, 1].

A paradigmatic and intriguing effect based on these empirical findings is visual search asymmetry: searching for an object, A, amidst other objects, B, can be substantially easier than searching for object B amongst instances of A. For example, detecting a curved line among straight lines is faster than searching for a straight line among curved lines. Search asymmetry has been observed in a wide range of tasks [25, 10, 23, 24]. Despite extensive phenomenological characterization, the mechanisms underlying how neural representations guide visual search and lead to search asymmetry remain mysterious.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

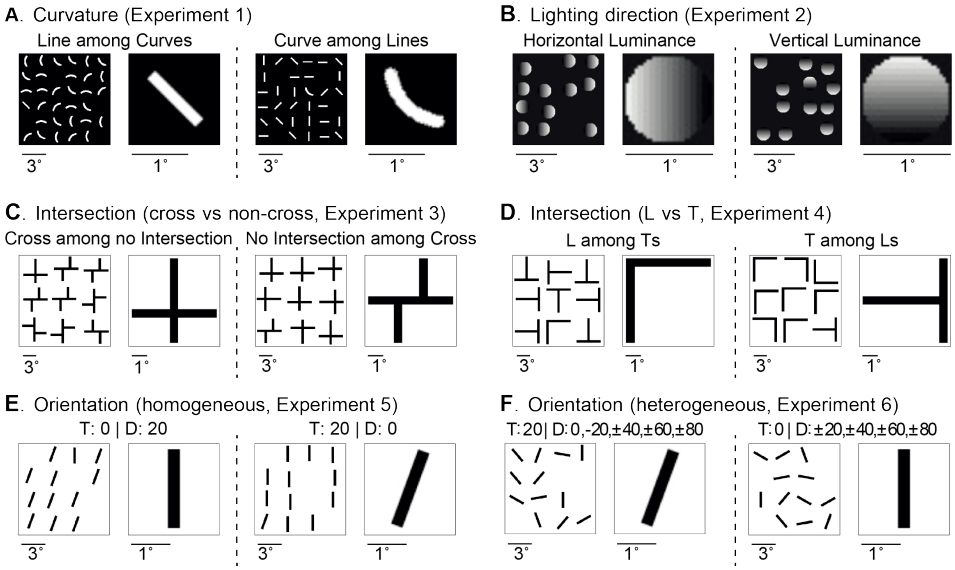

Figure 1: **Schematic illustration of the 6 experiments for asymmetry in visual search.** Each experiment has two conditions; in each condition, an example search image (left) and target image (right) are shown. The target image is smaller than the search image (see scale bars below each image). **A. Experiment 1.** Searching for a curve among lines ([25]). **B. Experiment 2.** Searching for vertical luminance changes among horizontal luminance changes ([10]). **C. Experiment 3.** Searching for shapes with no intersection among crosses ([23]). **D. Experiment 4.** Searching for rotated Ts among rotated Ls ([23]). **E. Experiment 5.** Searching for oblique lines among vertical lines ([24]). **F. Experiment 6.** Searching for vertical lines among oblique lines ([24]). In all cases, subjects find the target faster in the condition on the right.

We developed an image-computable model (eccNET) that aims to explain how visual representations are responsible for such search asymmetries and shed light on the fundamental inductive biases inherent to neural computations. Inspired by [27], the proposed visual search model combines eccentricity-dependent sampling, and top-down modulation through target-dependent attention. In contrast to deep convolutional neural networks (CNNs) that assume uniform resolution over all spatial locations, we introduce eccentricity-dependent pooling layers in the visual recognition processor, mimicking the separation between fovea and periphery in the primate visual system. As the task-dependent target information is essential during visual search [5, 1, 26], the model also incorporates top-down attention modulation and generates a sequence of eye movements until the target is found.

We examined six foundational psychophysics experiments showing visual search asymmetry and tested eccNET on the same stimuli. Importantly, the model was pre-trained for object classification on ImageNet and was not trained with the target and search images or with any human visual search data. The model spontaneously revealed strong asymmetries in visual search and qualitatively matched the polarity of human asymmetric behavior. We tested the hypothesis that asymmetry arises from the natural statistics seen during object classification tasks. We trained the eccNET model from scratch on an augmented version of ImageNet where we deliberately altered the stimulus statistics and re-tested the model on the search experiments. The polarity of search asymmetry disappeared or was altered according to the changes introduced in ImageNet, demonstrating that inductive biases in the developmental training set can govern complex visual behaviors.

2 Psychophysics Experiments in Visual Search Asymmetry

We examined 6 psychophysics experiments on visual search asymmetry [25, 10, 23, 24] (**Figure 1**). Subjects had to search for a given target randomly intermixed with distractor objects in a search array (see **Appendix A** for stimulus details). There were target-present and target-absent trials within random presentation sequences, and subjects had to press one of two keys to indicate that they found the target or that the target was absent.

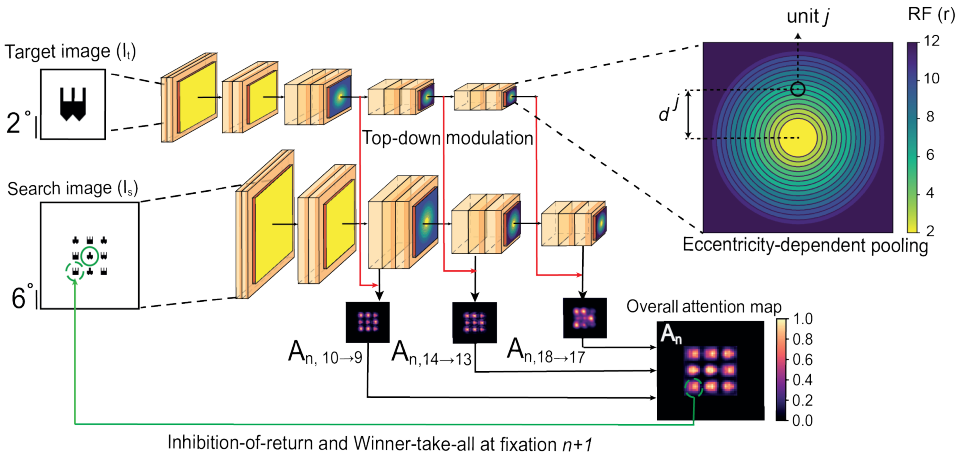

Figure 2: **Schematic of the computational model of visual search.** The model takes as input a target image (I_t) and a search image (I_s), both of which are processed through the same 2D-CNN with shared weights. At each fixation n, the model produces a top-down modulation map (A_n) that directs the next eye movement (Sec 3). The color bar on the right denotes attention values. Instead of using a typical 2D-CNN with uniform pooling window sizes, here we introduce an eccentricity-dependent sampling within each pooling layer of the network by modifying the VGG16 ([21]) architecture. We also provide illustration of eccentricity-dependent pooling layer l. It shows the receptive field sizes ($r^j_{l,n}$) for each unit j with distance d^j from the centre (the color bar here denotes the size of the pooling window in pixels).

Experiment 1: Curvature involved two conditions (**Figure 1A**) [25]: searching for (a) a straight line among curved lines, and (b) a curved line among straight lines. The target and distractors were presented in any of the four orientations: -45, 0, 45, and 90 degrees.

Experiment 2: Lighting Direction involved two conditions (**Figure 1B**) [10]: searching for (a) left-right luminance changes among right-left luminance changes, and (b) top-down luminance changes among down-top luminance changes. The luminance changes of each circle were discretized into 17 intensity levels.

Experiments 3-4: Intersection involved four conditions (**Figure 1C-D**) [23]: searching for (a) a cross among non-crosses, (b) a non-cross among crosses, (c) an L among Ts, and (d) a T among Ls. The target and distractors were presented in any of the four orientations: 0, 90, 180, and 270 degrees.

Experiments 5-6: Orientation involved four conditions (**Figure 1E-F**) [24]: searching for (a) a vertical line among 20-degrees-tilted lines, (b) a 20-degree-tilted line among vertical straight lines, (c) a 20-degree tilted line among tilted lines of from -80 to 80 degrees, and (d) a vertical straight line among tilted lines from -80 to 80 degrees.

3 Eccentricity-dependent network (eccNET) model

A schematic of the proposed visual search model is shown in **Figure 2**. All the source code and stimuli will be publicly available upon publication.

EccNet takes two inputs: a target image (I_t, object to search) and a search image (I_s, where the target object is embedded amidst distractors). EccNet starts fixating on the center of I_s and produces a sequence of fixations. EccNet uses a pre-trained 2D-CNN as a proxy for the ventral visual cortex to extract eccentricity-dependent visual features from I_t and I_s. At each fixation n, these features are used to calculate a top-down attention map (A_n). A winner-take-all mechanism selects the maximum of the attention map A_n as the location for the $n+1$-th fixation. This process iterates until EccNet finds the target with a total of N fixations. EccNet has infinite inhibition of return and therefore does not revisit previous locations. Humans do *not* have perfect memory and do re-visit previously fixated locations [28]. However, the total number of fixations in the six psychophysics experiments is small; therefore, the probability of having return fixations is small and the effect of having limited working memory capacity (limited inhibition of return) to store past fixations is negligible.

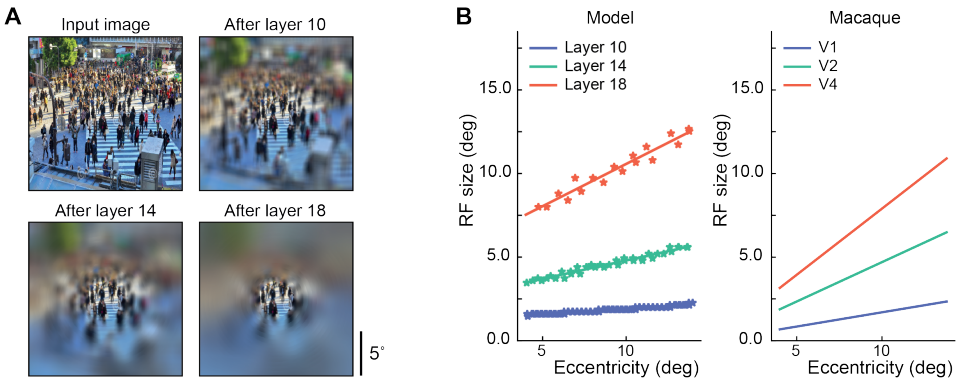

Figure 3: **Eccentricity dependence in the model matches that in macaque monkey visual cortex. A.** Visualization of one example image and its corresponding eccentricity-dependent sampling at pooling layers 3 through 5. **B.** Eccentricity-dependent sampling leads to increasing receptive field sizes as a function of eccentricity, in addition to increased receptive field sizes across layers for the model (left) and also for the macaque visual cortex ([6]).

EccNet needs to verify whether the target is present at each fixation. Since we focus here on visual search (target localization instead of recognition), we simplify the problem by bypassing this step and using an "oracle" recognition system. The oracle checks whether the selected fixation falls within the ground truth target location, defined as the bounding box of the target object. We only consider target-present trials for the model and therefore EccNet will always find the target.

Compared with the invariant visual search network (IVSN, [27]), we highlight novel components in EccNet, which we show to be critical in Sec 4:

1. Standard 2D-CNNs, such as VGG16 ([21]), have uniform receptive field sizes (convolution kernel sizes) within each layer. In stark contrast, visual cortex shows strong eccentricity-dependent receptive field sizes. Here we introduced eccentricity-dependent pooling layers, replacing all the max-pooling layers in VGG16.
2. Visual search models require computation of an overall attention map in order to decide where to fixate next. The attention map in IVSN was computed only once and did not change from one fixation to the next. Because the visual cortex module in eccNET changes with the fixation location in an eccentricity-dependent manner, here the attentional maps (A_n) are updated at each fixation n.
3. In contrast to IVSN where the top-down modulation happens only in a single layer, EccNet combines top-down modulated features across multiple layers.

Eccentricity-dependent pooling in visual cortex in EccNet. Receptive field sizes in visual cortex increase from one brain area to the next along the visual hierarchy (**Figure 3B**, right). This increase is captured by current computational models of visual recognition through pooling operations. In addition, receptive field sizes also increase with eccentricity *within* a given visual area (summarized in [6]; **Figure 3B**, right). Current 2D-CNNs assume uniform sampling within a layer and therefore do not reflect this eccentricity dependence. Here, we introduced eccentricity-dependent pooling layers in EccNet. We first define notations used in standard average pooling layers [7]. For simplicity, we describe the model components using pixels as units and the **Appendix D** shows scaling factors used to convert pixels to degrees of visual angles (dva) to compare with human behavior in Sec 2.

Traditionally, unit j in layer $l+1$ of VGG16 takes the average of all input units i in the previous layer l within its local receptive field of size r_{l+1} and its activation value y is given by: $y^j_{l+1} = \frac{1}{r_{l+1}} \sum_{i=0}^{r_{l+1}} y^i_l$

In the eccentricity-dependent operation, the receptive field size $r^j_{l+1,n}$ of input unit j in layer $l+1$ is a linear function of the Euclidean distance $d^j_{l+1,n}$ between input unit j and the current fixation location (fixation number n) on layer $l+1$.

$$r^j_{l+1,n} = \begin{cases} \lfloor \eta_{l+1}\gamma_{l+1}(d^j_{l+1,n}/\eta_{l+1} - \delta) + 2.5 \rfloor, & \text{if } d^j_{l+1,n}/\eta_{l+1} > \delta \\ 2, & \text{if } d^j_{l+1,n}/\eta_{l+1} < \delta \end{cases} \tag{1}$$

The floor function $\lfloor \cdot \rfloor$ rounds down the window sizes to the nearest integers. The variable γ_{l+1} is a positive scaling factor for layer $l+1$, defining how fast the receptive field size of unit j expands with respect to its distance from the fixation at layer $l+1$. The further away the unit j is from the current fixation, the larger the receptive field size (**Figure 2B**; in this figure, the fixation location happens to be at the center of the image). Therefore, the resolution is highest in the fixation location and decreases in peripheral regions. Based on the slope of eccentricity versus receptive field size in the macaque visual cortex ([6]), we experimentally set $\gamma_3 = 0.00$, $\gamma_6 = 0.00$, $\gamma_{10} = 0.14$, $\gamma_{14} = 0.32$, and $\gamma_{18} = 0.64$ (see **Figure 3B** for the slopes of eccentricity versus receptive field sizes over pooling layers. We define δ as the constant fovea size. For those units within the fovea, we set a constant receptive field size of 2 pixels. Constant η_{l+1} is a positive scaling factor which converts the dva of the input image to the pixel units at the layer l (see **Appendix D** for specific values of η_{l+1}). As in the stride size in the original pooling layers of VGG16, we empirically set a constant stride of 2 pixels for all eccentricity-dependent pooling layers.

A visualization for $r^j_{l,n}$ is shown in **Figure 2**, where different colors denote how the receptive field sizes expand within a pooling layer. Local receptive field sizes $r^j_{l,n}$ versus eccentricity $d^j_{l,n}$ in pixels at each pooling layer are reported in the **Table S1**. **Figure 3A** illustrates the change in acuity at different pooling layers. These customized eccentricity-dependent pooling layers can be easily integrated into other object recognition deep neural networks. All the computational steps are differentiable and can be trained end-to-end with other layers. However, since our goal is to test the generalization of EccNet from object recognition to visual search, we did not retrain EccNet and instead, we used the pre-trained weights of VGG16 on the ImageNet classification task.

Top-down attention modulation across multiple layers of visual cortex in EccNet. Given the current n^{th} fixation location, the visual cortex of EccNet with eccentricity-dependent pooling layers extracts feature maps $\phi^t_{l+1,n}$ at layer $l+1$ for the target image I_t. Correspondingly, the model extracts $\phi^s_{l,n}$ in response to the search image I_s. We define the top-down modulation map $A_{l+1\to l,n}$ as:

$$A_{l+1\to l,n} = m(\phi^t_{l+1,n}, \phi^s_{l,n}) \tag{2}$$

where $m(\cdot)$ is the target modulation function defined as a 2D convolution with $\phi^t_{l+1,n}$ as convolution kernel operating on $\phi^s_{l,n}$. Note that the layer $l+1$ modulates the activity of layer l. We empirically selected $l = 9, 13, 17$ as the layers where top-down modulation is performed.

To compute the overall top-down modulation map A_n, we first resize $A_{10\to 9,n}$ and $A_{14\to 13,n}$ to be of the same size as $A_{18\to 17,n}$. EccNet then takes the weighted linear combination of normalized top-down modulation maps across all three layers.

$$A_n = \sum_{l=9,13,17} w_{l,n} \frac{(A_{l+1\to l,n} - \min A_{l+1\to l,n})}{(\max A_{l+1\to l,n} - \min A_{l+1\to l,n})} \tag{3}$$

where $w_{l,n}$ are weight factors governing how strong top-down modulation at the lth layer contributes to the overall attention map A_n. These weights $w_{l,n}$ were calculated during the individual search trials using the maximum activation value obtained from each individual top-down modulation map:

$$w_{l,n} = \frac{\max A_{l+1\to l,n}}{\sum_{i=9,13,17} \max A_{i+1\to i,n}} \tag{4}$$

Comparison with human performance and baseline models. The psychophysics experiments in Sec 2 did not measure eye movements and instead report a *key press reaction time* (RT) in milliseconds whereby subjects presumably found the target. To compare fixation sequences predicted by EccNet with reaction times measured in human experiments, we conducted a similar experiment as Experiment 4 in Sec 2 using eye tracking (**Appendix C**). Since RT results from a combination of time taken by eye movements plus a finger motor response time, we performed a linear least-squares regression (**Appendix B**) on the eye tracking and RT data collected from this experiment by fitting RT as a function of number of fixations N until the target was found: $RT = \alpha * N + \beta$. The slope $\alpha = 252.36$ ms/fixation approximates the duration of a single saccade and subsequent fixation, and the intercept $\beta = 376.27$ can be interpreted as a finger motor response time. We assumed that α and β are approximately independent of the task used the same values to convert the number of fixations predicted by EccNet to RT throughout all 6 experiments.

We introduce three evaluation metrics. In each (Sec 2), we evaluated the model and human performance showing **key press reaction times (RT)** as a function of number of items on the stimulus, as commonly used in psychology [25, 10, 23, 24]. We computed the search slope of RT versus number of items per condition in each experiment. We then define a new metric called **Asymmetry Index** as the difference in slopes between the two conditions divided by the sum of the slopes. Here the difference is taken such that the Asymmetry Index for humans becomes positive. [Mengmi: I found the english can improve... check lanuguage, grammar, etc throughout the entire text] Thus, if for a given experiment if the model follows the asymmetry pattern of the humans than it will give a positive Asymmetry Index otherwise a negative Asymmetry Index. We then compare two models by taking an overall average of all the six Asymmetry Indexes and by comparing the total number experiment which is giving positive Asymmetry Index. Higher the average value or the higher the number of points indicates better modelling of asymmetry pattern to humans.

We included the following baselines for comparison with EccNet:

Chance: a sequence of fixations is generated by uniform random sampling.

pixelMatching: the attention map is generated by sliding the raw pixels of I_t over I_s (stride $= 1 \times 1$).

GBVS [9]: we used bottom-up saliency as the attention map for generating fixations.

IVSN [27]: the top-down attention map is based on the features from the top layer of VGG16.

4 Results

4.1 EccNet qualitatively captures visual search asymmetry

Figure 4A (left plot, red line) shows the results of Experiment 1 (**Figure 1**, left), where subjects looked for a straight line amongst curved lines. Increasing the number of distractors led to longer reaction times, a classical observation in visual search studies. When the target and distractors were reversed and subjects had to search for a curved line in the midst of straight lines, the reaction times were lower and showed minimal dependence on the number of objects (**Figure 4A**, left, blue line). In other words, it is easier to search for a curved line amidst straight lines than the reverse.

The same target and search images were presented to EccNet. Example fixation sequences from the model for Experiment 1 are shown in **Figure S10A**. **Figure 4A** (right) shows EccNet's reaction time as a function of the number of objects for Experiment 1. EccNet had not seen any of these types of images before and the model was never trained to search for curves or straight lines. Critically, there was no explicit training to show an asymmetry between the two experimental conditions. Remarkably, EccNet qualitatively captured the key observations from the psychophysics experiments: When searching for curves among straight lines (blue), the reaction times were largely independent of the number of distractors. In contrast, when searching for straight lines among curves (red), the reaction times increased substantially with the number of distractors. The model's absolute reaction times were far from identical from human RTs. Of note, there was no training to quantitatively fit the human data (Sec. 5). In sum, without any explicit training, EccNet qualitatively captures this fundamental asymmetry in visual search behavior.

We considered five additional experiments (Sec. 2) that reveal conceptually similar asymmetries, but which are based on searching for a wide range of distinct features (**Figure 1B-F**). In Experiment 2, it is easier to search for top-down luminance changes than left-right luminance changes (**Figure 1B** [10]). In Experiments 3-4, it is easier to find non-intersections among crosses than the reverse (**Figure 1C**; [23]), and it is easier to find a a rotated letter T among rotated letters L than the reverse (**Figure 1D**; [23]). In Experiments 5-6, subjects can find a tilted bar amongst vertical distractors more efficiently than the reverse when distractors are of homogeneous type (**Figure 1E**; [24]) but the situation reverses when the distractors are of heterogenous type (**Figure 1E-F**; [24]). In all of these cases, the psychophysics results reveal lower reaction times and only a weak increase in reaction time with increasing numbers of distractors for the easier search condition and a more pronounced increase in reaction time with more distractors for the harder condition (**Figure 4A-F**, left panels).

Without any image-specific or task-specific training, EccNet qualitatively captured these asymmetries (**Figure 4A-D, F**, right panels). As noted in Experiment 1, EccNet did not necessarily match the human reaction times at a quantitative level. While in some cases the reaction times for EccNet were comparable to the human ones (e.g., **Figure 4C**), in other cases, there were large differences

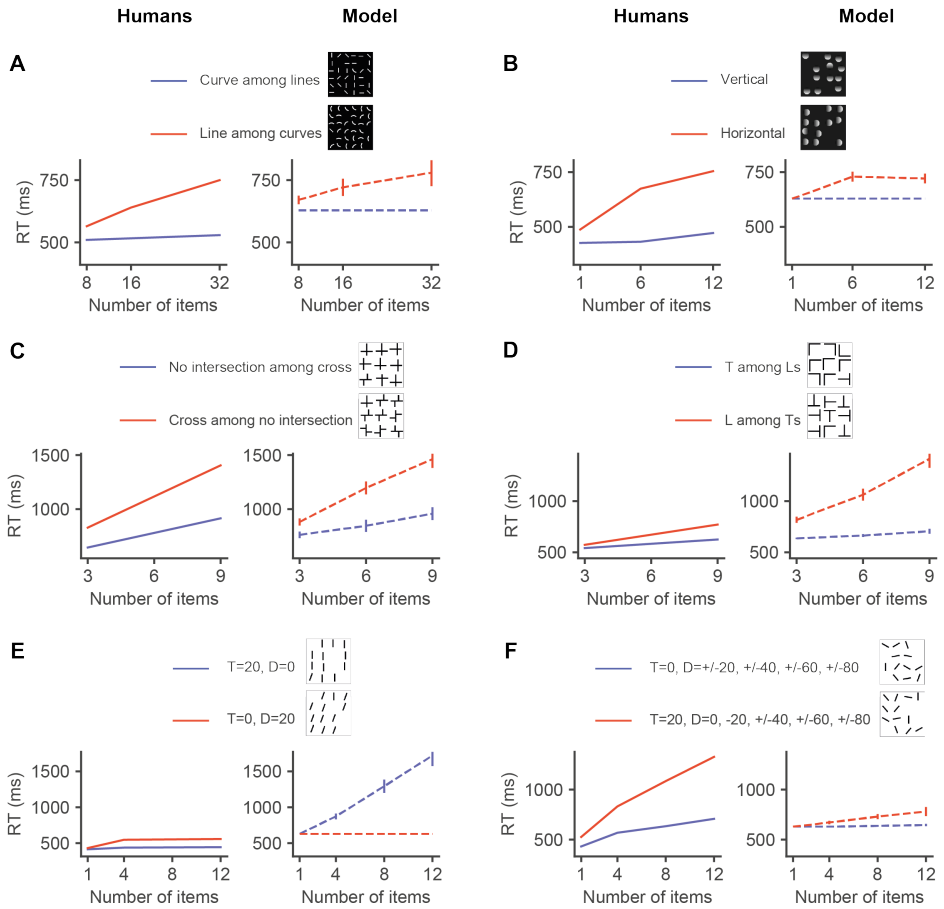

Figure 4: **The model shows visual search asymmetry and qualitatively captures human behavior.** Reaction time as a function of the number of items in the display for each of the six experiments in **Figure 1** for humans (left, solid lines) and for the model (right, dashed lines). The results for humans were obtained from the references in **Figure 1**. The line colors denote the different experimental conditions. Error bars denote standard error (SE).

(e.g., **Figure 4D**). Despite the shifts along the y-axis in terms of the absolute reaction times, EccNet qualitatively reproduced the visual search asymmetries in majority of experiments. However, in Experiment 5 where humans showed minimal search asymmetry effect, EccNet fails to show same polarity of asymmetry search as humans and takes longer RT than humans when finding tilted lines amongst vertical lines (blue dash, see Sec 5).

As the slope of RT have been commonly used to evaluate human search efficiency, to summarize and directly compare the human and model results, we computed the slope of RT versus number of object plots and calculated the Asymmetry Index (see Sec 3) in **Figures 4**. A linear fit was performed for each condition in each experiment. The slope, in units of milliseconds per object, reflects the additional cost incurred when adding more distractors. A small slope reflects easier search conditions. **Figure 5A** compares the Asymmetry Indexes for each condition in humans versus the model. There was general agreement between the Asymmetry Indexes predicted by the model and those measured in the human psychophysics experiments (**Figure 5A**). Conditions that shows positive Asymmetry Index for humans tended to be positive for the model. Of note, since Asymmetry Index are calculated using the slopes, they are independent of shifts along the y-axis in **Figure 4**. Thus, the model Asymmetry Indexes approximate the psychophysics Asymmetry Indexes even in cases where there is poor agreement in the absolute RT values.

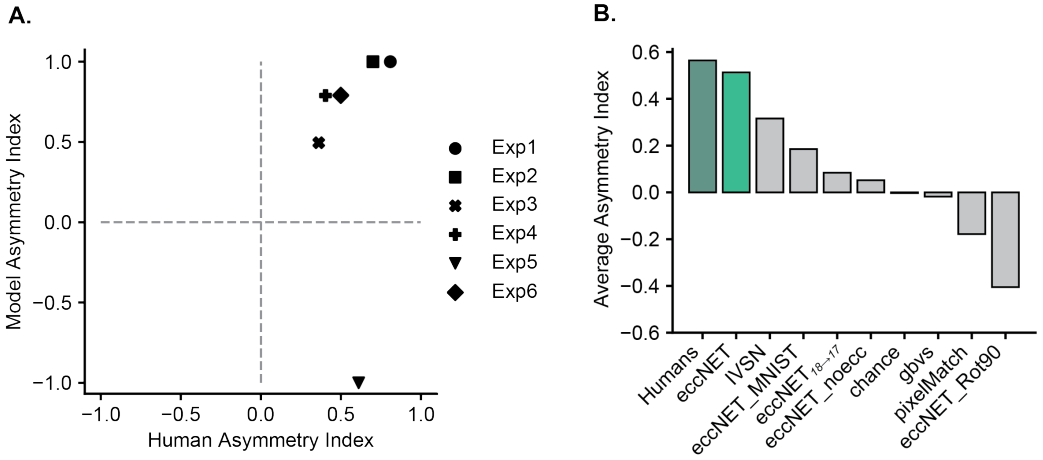

Figure 5: **The model approximates the human Asymmetry Indexes calculated using search cost slopes. A** shows the Asymmetry Indexes (see Sec 3) for the model (y-axis) versus humans (x-axis) for all six experiments. The gray dashed lines indicates the lines passing through the origin. Points lieing in the 1st quadrant indicates that the model follows the asymmetry pattern of humans . **B. Comparison to ablated and alternative models.** Average of the Asymmetry Indexes across all the six experiments for humans and different models. The dark green bar corresponds to humans and the light green bar corresponds to our proposed eccNET model. The different models are defined in Sec 3.

4.2 Baseline models fail to capture search asymmetry

We compared the performance of EccNet with several baseline models **Figure 5B** (see also **Figure S2, 6, 8, 9** for analysis of baseline model performances in terms of RT versus number of items and Asymmetry Indexes). All the baseline models also had infinite inhibition of return, oracle recognition, and the same conversion of fixations into RT values.

As a null hypothesis, we considered a model where fixations landed on random locations. This model consistently required longer reaction times and yielded an Asymmetry Index value close to 0 that showed no-correlation with human asymmetry property (**Figure 5B**, *chance*). A simple template-matching algorithm also failed to capture human performance (**Figure 5B**, *pixelMatch*), showing that features at pixel level are insufficient to explain human visual search. Next, we considered a purely bottom-up algorithm that relied exclusively on saliency; the failure of this model shows that it is important to take into account the top-down target features to compute the overall attention map (**Figure 5B**, *gbvs*).

None of these baseline models above contain a dictionary of complex features from natural images. We reasoned that the previous IVSN architecture ([27]), which was exposed to almost a million natural images through ImageNet, would yield better performance. Indeed, IVSN showed a higher Asymmetry Indexes to humans than the other baselines, yet its performance was well below that of EccNet (**Figure 5B**, *IVSN*). This observation shows that the components introduced in EccNet play critical role and we next investigate those components by selectively ablating different model components.

4.3 Model ablations reveal essential components contributing to search asymmetry

We systematically ablated two essential components of EccNet (**Figure 5B**). In EccNet, top-down modulations occurs at three levels: layer 10 to layer 9 ($A_{10\rightarrow 9}$), 14 to 13 ($A_{14\rightarrow 13}$), and 18 to 17 ($A_{18\rightarrow 17}$). We considered EccNet$_{18\rightarrow 17}$, where top-down modulation only happened at layer 17 of the VGG16 architecture. EccNet$_{18\rightarrow 17}$ yielded a lower average of Asymmetry Indexes (0.084), suggesting that visual search benefits from the combination of layers for top-down attention modulation.

To model the distinction between foveal and peripheral vision, we introduced eccentricity-dependent pooling layers. The resulting increase in receptive field size is qualitatively consistent with the receptive field sizes of neurons in the macaque visual cortex (**Figure 3B**). To assess the impact of this step, we removed all eccentricity-dependent pooling layers (EccNet$_{noecc}$). The lower average of

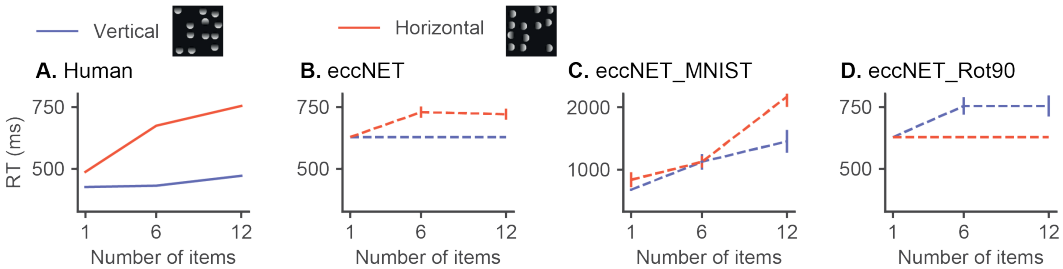

Figure 6: **Training data biases polarity of search asymmetry.** Reaction times as a function of the number of items for Experiment 2 on illumination conditions. (**A-B**) Humans and EccNet (same results as in **Figure 4B** reproduced here for comparison purposes). (**C**) EccNet trained on the MNIST [4]. (**D**) EccNet trained on ImageNet [12] after 90 degree rotation.

Asymmetry Indexes implies that eccentricity-dependent pooling is an essential component in EccNet for asymmetry in visual search.

4.4 Training data biases polarity of search asymmetry

Having an image-computable model enables us to dissect the mechanisms responsible for visual search asymmetry. Given that the model was not designed, or excplicitly trained to achieve such asymmetries, we were intrigued by how asymmetry could emerge. We hypothesized that these asymmetry patterns can be acquired from the visual features learnt through the natural statistics of the training images in EccNet. To test this hypothesis, as a proof-of-principle, we focused on Experiment 2 (**Figure 1B**).

We trained the visual cortex part of EccNet from scratch using a training set with completely different image statistics. Instead of using ImageNet containing millions of natural images, we trained EccNet on the MNIST dataset [4], which contains binary images of hand-written digits. We tested EccNet_MNIST in Experiment 2. Indeed, the asymmetry effect of EccNet_MNIST disappeared and the absolute reaction times required to find the target increased substantially (**Figure 6C**). The overall average of Asymmetry Indexes also significantly reduced (**Figure 5B**).

These observations point to the critical role of the training set. To better understand which aspects of natural image statistics are critical for asymmetry to emerge, we focused on the distinction between different light directions. We conjectured that humans are more used to light coming from a vertical than a horizontal directions and that this bias would be present in the ImageNet dataset. To test this idea, we trained EccNet from scratch using an altered version of the training set from ImageNet, where all the training images are rotated by 90 degrees counter-clockwise; and then tested EccNet on the same stimuli in Experiment 2. Note that random image rotation is not one of the original data augmentation steps for the VGG16 model pre-trained on ImageNet [21], thus, the VGG16 pre-trained on ImageNet has only been exposed to vertical lighting conditions. Surprisingly, the results of EccNet trained on the 90-degree-rotated ImageNet (EccNet_Rot90) show that the polarity of asymmetry is reversed (**Figure 6**) compared with humans, and the absolute RT difference of EccNet_Rot90 between two search conditions remains roughly the same as EccNet. To verify that the features learnt from ImageNet with training images rotated 90 degrees are preserved for recognition, we tested the visual cortex of EccNet_Rot90 on the original test set of ImageNet and obtained a top-1 image classification accuracy of 66.7% which is similar to the original result for VGG16 pre-trained on ImageNet (top-1 accuracy of 72%). These observations show that the features learnt during object recognition are useful for guiding visual search and the statistics from the training images contribute to visual search asymmetry.

5 Discussion

We examined six classical and paradigmatic experiments demonstrating a strong search asymmetry. Humans find a target, A, amidst distractors, B, much faster than in the reverse search condition of B among A. Given the similarity between the target and distractors, it is not clear why one condition should be easier than the other. Therefore, these asymmetries reflect strong priors in how the visual system processes features to guide search behavior. We propose EccNet, a deep convolutional neural

network architecture introducing eccentricity-dependent sampling and top-down attention modulation. At the heart of the model is a "ventral visual cortex" module pre-trained in a classification task on ImageNet [17]. The model had no previous exposure to the images in the current study, which are not in ImageNet. Moreover, the model *was not trained in any of these tasks*, did not have any tuning parameters dependent on eye movement or RT data, and was not designed with the goal of showing asymmetry. Strikingly, despite this lack of tuning or training, asymmetric search behavior emerges from the model. Furthermore, EccNet captures observations in five out of six human psychophysics experiments in terms of the search cost and polarity of asymmetry across different conditions within each experiment and Asymmetry Indexes close to humans across experiments. It is possible that EccNet fails to capture asymmetry in Experiment 5 because EccNet does not incorporate bottom-up saliency effect where 20 degrees titled lines gets more easily "popped out" than vertical lines.

An image-computable model allows us to examine potential mechanisms underlying asymmetric search. Even though the target and search images are different from those in ImageNet, natural images do contain edges of different size, color, and orientation. To the extent that the distribution of image statistics in ImageNet reflects the natural world, a question that has been contested and deserves further scrutiny, one might expect that the training set could capture some of the priors inherent to human perception. We considered this conjecture further in the case of Experiment 2. The asymmetry in vertical versus horizontal illumination changes has been attributed to the fact that animals are more used to seeing light coming from the top (the sun), a bias that is also likely to be reflected in ImageNet. Consistent with this idea, we demonstrate that changing the training diet for the model alters its visual search behavior. Specifically, rotating the images by 90 degrees, leading to a change in the illumination direction used to train the visual cortex, leads to a reversal in the polarity.

Although the model *qualitatively* captures the critical observations in the psychophysics experiments, the model does not always yield accurate estimates of the absolute reaction times. Several factors might contribute towards the discrepancy between the model and humans. First and foremost, EccNet lacks multiple critical components of human visual search abilities including finite working memory, object recognition, and contextual reasoning, among others. Second, to convert fixations from EccNet to key press reaction time, we assume that the relationship between number of fixations and RTs is independent of the experimental conditions. This assumption probably constitutes an oversimplification. While the approximate frequency of saccades tends to be similar across tasks, there can still be differences in the saccade frequency and fixation duration depending on the nature of the stimuli, on the difficulty of the task, on some of the implementation aspects such as monitor size and contrast, and even on the subjects themselves. Third, the human psychophysics tasks involved both target present and target absent trials. Deciding whether an image contains a target or not (human psychophysics experiments) is not identical to generating a sequence of saccades until the target is fixated upon (model). It is likely that humans located the target when they pressed a key to indicate target present, but this was not measured in the experiments and it is conceivable that a key was pressed while the subject was still fixating on a distractor. Fourth, humans tend to make "return fixations", whereby they fixate on the target, move the eyes away and come back to the target [28] while the current model would stop at the first fixation on the target.

Overall, we introduce a biologically plausible visual search model that can approximate a fundamental property of human visual behavior. The success of our model encourages further investigation of improved computational models of visual search and emphasizes the importance of directly comparing models with biological architectures and behavioral outputs.

References

[1] Narcisse P Bichot, Matthew T Heard, Ellen M DeGennaro, and Robert Desimone. A source for feature-based attention in the prefrontal cortex. *Neuron*, 88(4):832–844, 2015.

[2] James W Bisley. The neural basis of visual attention. *The Journal of physiology*, 589(1):49–57, 2011.

[3] Maurizio Corbetta. Frontoparietal cortical networks for directing attention and the eye to visual locations: identical, independent, or overlapping neural systems? *Proceedings of the National Academy of Sciences*, 95(3):831–838, 1998.

[4] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[5] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

[6] Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011.

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[8] Marie-Hélène Grosbras, Angela R Laird, and Tomás Paus. Cortical regions involved in eye movements, shifts of attention, and gaze perception. *Human brain mapping*, 25(1):140–154, 2005.

[9] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19:545–552, 2006.

[10] Dorothy A Kleffner and Vilayanur S Ramachandran. On the perception of shape from shading. *Perception & Psychophysics*, 52(1):18–36, 1992.

[11] Gabriel Kreiman and Thomas Serre. Beyond the feedforward sweep: feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, 1464(1):222, 2020.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[13] Jacob L Orquin and Simone Mueller Loose. Attention and choice: A review on eye movements in decision making. *Acta psychologica*, 144(1):190–206, 2013.

[14] Sofia Paneri and Georgia G Gregoriou. Top-down control of visual attention by the prefrontal cortex. functional specialization and long-range interactions. *Frontiers in neuroscience*, 11:545, 2017.

[15] John H Reynolds and Leonardo Chelazzi. Attentional modulation of visual processing. *Annu. Rev. Neurosci.*, 27:611–647, 2004.

[16] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[18] Alexander C Schütz, Doris I Braun, and Karl R Gegenfurtner. Eye movements and perception: A selective review. *Journal of vision*, 11(5):9–9, 2011.

[19] T Serre, A Oliva, and T Poggio. Feedforward theories of visual cortex account for human performance in rapid categorization. *PNAS*, 104(15):6424–6429, 2007.

[20] Thomas Serre, Gabriel Kreiman, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, and Tomaso Poggio. A quantitative theory of immediate visual recognition. *Progress in brain research*, 165:33–56, 2007.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Sabine Kastner Ungerleider and Leslie G. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 23(1):315–341, 2000.

[23] Jeremy M Wolfe and Jennifer S DiMase. Do intersections serve as basic features in visual search? *Perception*, 32(6):645–656, 2003.

[24] Jeremy M Wolfe, Stacia R Friedman-Hill, Marion I Stewart, and Kathleen M O'Connell. The role of categorization in visual search for orientation. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1):34, 1992.

[25] Jeremy M Wolfe, Alice Yee, and Stacia R Friedman-Hill. Curvature is a basic feature for visual search tasks. *Perception*, 21(4):465–480, 1992.

[26] JeremyM Wolfe and ToddS Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):1–8, 2017.

[27] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, 9(1):1–15, 2018.

[28] Mengmi Zhang, Will Xiao, Olivia Rose, Katarina Bendtz, Margaret Livingstone, Carlos Ponce, and Gabriel Kreiman. Look twice: A computational model of return fixations across tasks and species. *arXiv preprint arXiv:2101.01611*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sec. 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We submitted as a zip file together with supplementary material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix G
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For all the applicable figures, we provided SE (standard errors) as errorbars.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix G
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Together with source code, we uploaded all the stimuli in zip file.
   (d) Did/ you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix C and FigS1
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Appendix C