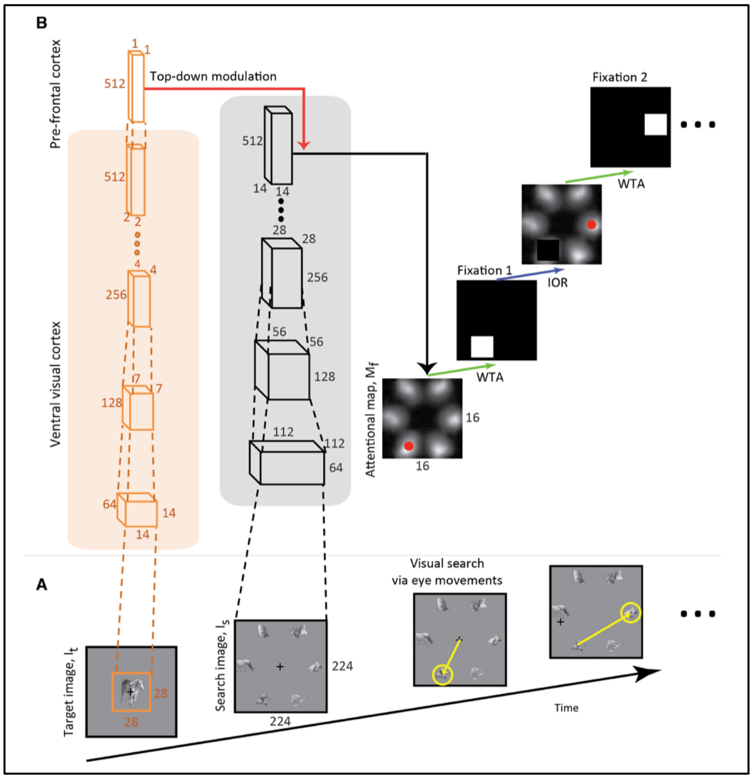
Finding any Waldo: zero-shot invariant and efficient visual search

Mengmi Zhang and Gabriel Kreiman

Visual search constitutes a ubiquitous challenge in natural vision, including daily tasks such as finding a friend in a crowd or searching for a car in a parking lot. Visual search must fulfill four key properties: selectivity (to distinguish the target from distractors in a cluttered scene), invariance (to localize the target despite changes in its rotation, scale, illumination, and even searching for generic object categories), speed (to efficiently localize the target without exhaustive sampling), and generalization (to search for any object, even ones that we have had minimal or no experience with). Here we propose a computational model that is directly inspired by neurophysiological recordings during visual search in macaque monkeys, which maps the discriminative power from object recognition models to the problem of visual search. The model takes two inputs, a target object, and a search image, and produces a sequence of fixations. The model consists of a deep convolutional network that extracts features about the target object, stores those features, and uses those features in a top-down fashion to modulate the responses to the search image, thus generating a task-dependent saliency map. We show that the model fulfills the critical properties outlined above, distinguishing it from heuristic approaches such as template matching, random search, sliding windows, bottom-up saliency maps and object detection algorithms. Furthermore, we directly compare the model against human eye movement behavior during three increasingly more complex tasks where subjects have to search for a target object in a multi-object array image, in natural scenes or in the well-known Waldo search task. We show that the model provides a reasonable first-order approximation to human behavior and can efficiently find targets in an invariant manner, without any training for the target objects.

Model schematic. A. *Sequence of events during the visual search task. A target image is presented, followed by a search image where subjects move their eyes to locate the target object.* ***B***. *Architecture of the model, referred to as Invariant Visual Search Network (IVSN). The model consists of a pre-trained 16-layer bottom-up hierarchical network (VGG-16) that mimics image processing in the ventral visual cortex. Only some of the layers are shown here for simplicity, the dimensions of the feature maps are indicated for each layer. The model generates features in each layer when presented with the target image I_t. The top-level features are stored in a pre-frontal cortex module that contains the task-dependent information about the target in each trial. Top-down information from pre-frontal cortex modulates the features obtained in response to the search image, I_s, by convolving the target presentation of I_t wit the top-level feature map from I_s, generating the attention map M_f. A winner-take-all mechanism (WTA, green) chooses the maximum in the attention map (red dot) as the location for the next fixation. If the target is not found at the current fixation, inhibition of return (IOR, blue), the fixation location is set to 0 in the attention map and the next maximum is selected. This process is repeated until the target is found.*


B
Pre-frontal cortex
Top-down modulation
1
1
512
512
2
2
4
4
256
7
7
128
64
14
14
Ventral visual cortex
512
14
14
28
28
256
56
56
128
112
112
64
Attentional map, M_f
16
16
WTA
Fixation 1
IOR
WTA
Fixation 2
A
Target image, I_t
28
28
Search image, I_s
224
224
Visual search via eye movements
Time