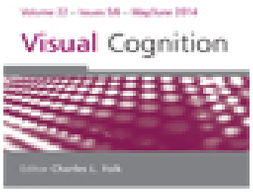
Routledge
Taylor & Francis Group



Visual Cognition

ISSN: 1350-6285 (Print) 1464-0716 (Online) Journal homepage: https://www.tandfonline.com/loi/pvis20

Modelling attention control using a convolutional neural network designed after the ventral visual pathway

Chen-Ping Yu, Huidong Liu, Dimitrios Samaras & Gregory J. Zelinsky

To cite this article: Chen-Ping Yu, Huidong Liu, Dimitrios Samaras & Gregory J. Zelinsky (2019): Modelling attention control using a convolutional neural network designed after the ventral visual pathway, Visual Cognition, DOI: 10.1080/13506285.2019.1661927

To link to this article: https://doi.org/10.1080/13506285.2019.1661927

VISUAL COGNITION
https://doi.org/10.1080/13506285.2019.1661927

Routledge
Taylor & Francis Group

Modelling attention control using a convolutional neural network designed after the ventral visual pathway

Chen-Ping Yu[a,c], Huidong Liu[a], Dimitrios Samaras[a] and Gregory J. Zelinsky[a,b]

[a]Department of Computer Science, Stony Brook University, Stony Brook, NY, USA; [b]Department of Psychology, Stony Brook University, Stony Brook, NY, USA; [c]Department of Psychology, Harvard University, Cambridge, MA, USA

ABSTRACT

We recently proposed that attention control uses object-category representations consisting of category-consistent features (CCFs), those features occurring frequently and consistently across a category's exemplars [Yu, C.-P., Maxfield, J. T., & Zelinsky, G. J. (2016). Searching for category-consistent features: A computational approach to understanding visual category representation. *Psychological Science*, *27*(6), 870–884.] Here we extracted from a Convolutional Neural Network (CNN) designed after the primate ventral stream (VsNet) CCFs for 68 object categories spanning a three-level category hierarchy, and evaluated VsNet against the gaze behaviour of people searching for the same categorical targets. We also compared its success in predicting attention control to two other CNNs that differed in their degree and type of brain inspiration. VsNet not only replicated previous reports of stronger attention guidance to subordinate-level targets, but with its powerful CNN-CCFs it predicted attention control to individual target categories. Moreover, VsNet outperformed the other CNN models tested, despite these models having more trainable convolutional filters. We conclude that CCFs extracted from a brain-inspired CNN can predict goal-directed attention control.

ARTICLE HISTORY
Received 4 March 2019
Accepted 12 August 2019

KEYWORDS
Brain-inspired CNN; attention control; categorical search; category-consistent features

The brain's ability to flexibly exert top-down control over motor behaviour is key to achieving visuomotor goals and performing everyday tasks (Ballard & Hayhoe, 2009), but a neurocomputational understanding of goal-directed attention control is still in its infancy. Here we introduce VsNet, a neurocomputational model inspired by the primate ventral stream of visually-responsive brain areas, that predicts attention control by learning the representative visual features of an object category.

VsNet advances existing models of attention control in several respects. First, it is image computable, meaning that it accepts the same visually complex and unlabelled imagery that floods continuously into the primate visual system (see also Adeli, Vitu, & Zelinsky, 2017; Zelinsky, Adeli, Peng, & Samaras, 2013). This is essential for a model aimed at understanding attention control in the real world, as objects do not come with labels telling us what and where they are. Note that although there are several excellent image-computable models of fixation prediction (Bylinskii, Judd, Oliva, Torralba, & Durand, 2016), these are all in the context of a free-viewing task and therefore outside of our focus on goal-specific attention control.[1] Second, VsNet is among the first uses of a convolutional neural network (CNN) to predict goal-directed attention. CNNs are one class of artificial deep neural networks that have been setting new performance benchmarks over diverse domains, not the least of which is the automated (without human input) recognition of visually-complex categories of objects (He, Zhang, Ren, & Sun, 2016; Krizhevsky, Sutskever, & Hinton, 2012; Russakovsky et al., 2015; Simonyan & Zisserman, 2015). However, CNN models that predict goal-directed attention control are still uncommon (Adeli & Zelinsky, 2018; Zhang et al., 2018). A third and core source of VsNet's capacity to predict attention control is its extraction of the visual features from image exemplars that are most representative of an object category. In short, VsNet harnesses the power of deep learning to extract the *category-consistent features* (Yu, Maxfield, & Zelinsky, 2016) used by the ventral visual areas to control the goal-directed application of attention.

CONTACT Gregory J. Zelinsky gregory.zelinsky@stonybrook.edu Psychology B-240, Stony Brook University, Stony Brook, NY 11794–2500, USA

VsNet is novel in that it is a brain-inspired CNN. There is the start of an interesting new discussion about how biological neural networks might inform the design of artificial deep neural networks (Grill-Spector, Weiner, Gomez, Stigliani, & Natu, 2018; Kietzmann, McClure, & Kriegeskorte, 2019), and VsNet is the newest addition to this discussion. Our approach is neurocomputational in that, given the many ways that CNNs can be built, we look to the rich neuroscience literature for design inspiration and parameter specification. Most broadly, VsNet is a multi-layered deep network, making its architecture analogous to the layers of brain structures existing along the ventral pathway. The brain's retinotopic application of filters throughout most of these ventral areas also embody a parallelized convolution similar to unit activation across a CNN's layers (Cadieu et al., 2014; Hong, Yamins, Majaj, & DiCarlo, 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). This parallel between a CNN and the ventral stream's organization has not gone unnoticed (Kriegeskorte, 2015), and unit activation across the layers of a CNN has even been used to predict neural activity recorded from brain areas in response to the same image content (Cadieu et al., 2014; Yamins et al., 2014). VsNet extends this work by making the architecture of its layers also brain-inspired, each modelled after a specific brain area in the primate ventral stream. In contrast, existing neurocomputational efforts have used either AlexNet (Krizhevsky et al., 2012) or one of its feed-forward variants (Simonyan & Zisserman, 2015; Szegedy, Liu, Jia, Sermanet, & Reed, 2015; Zeiler & Fergus, 2014), which are pre-trained CNNs designed purely to win image classification competitions (e.g., the ILSVRC2012 challenge, also known as ImageNet, Russakovsky et al., 2015) without regard for the structural and functional organization of the primate ventral visual system. The same disregard for neurobiological constraint applies to later generations of deep networks using different architectures (He et al., 2016; Huang, Liu, Van Der Maaten, & Weinberger, 2017; Zagoruyko & Komodakis, 2016). Determining how VsNet's performance compares to less brain-inspired CNNs is one broad aim of our study, with our hypothesis being that a model's predictive success will improve as its architecture becomes more like that of the primate brain.

A second broad aim is to predict people's goal-directed allocation of overt attention as they search for categories of objects. CNNs have been used to predict the bottom-up allocation of attention in scenes (Huang, Shen, Boix, & Zhao, 2015; Li & Yu, 2015; Wang & Shen, 2017), but they have only just started to be used to model the top-down control of attention (Adeli & Zelinsky, 2018; Zhang et al., 2018). We operationally define attention control as the degree that eye movements from human participants are guided to targets in a categorical search task. The spatial locations fixated via eye movements are an ideal behavioural ground truth for our purpose, as an eye movement is the most basic observable behaviour linked to a covert shift of spatial attention (Deubel & Schneider, 1996). Our focus on categorical search is similarly perfect. Categorical search, the search for an object designated only by its category name, can be contrasted with exemplar search, the more common task where participants are cued with an image showing the exact object that they are to search for. Categorical search therefore blends a highly nontrivial object classification task with a gold-standard measure of attention control, the oculomotor guidance of gaze to a target (Zelinsky, 2008).

While historically a neglected task for studying attention control (see Zelinsky, Peng, Berg, & Samaras, 2013, for discussion), interest in categorical search has accelerated in recent years (e.g., Cohen, Alvarez, Nakayama, & Konkle, 2016; Hout, Robbins, Godwin, Fitzsimmons, & Scarince, 2017; Nako, Wu, & Eimer, 2014; Peelen & Kastner, 2011), a growth fuelled by several key observations: (1) that attention *can* be guided to target categories, as exemplified by the above-chance direction of initial search saccades to target category exemplars in search arrays (Yang & Zelinsky, 2009), (2) that the strength of the control signal guiding attention to categorical targets depends on the amount of target-defining information provided in the category cue (e.g., stronger guidance for "work boot" than "footwear"; Schmidt & Zelinsky, 2009), (3) that search is guided to distractors that are visually similar to the target category (guidance to a hand fan when searching for a butterfly; Zelinsky, Peng, & Samaras, 2013), (4) that guidance improves with target typicality (stronger guidance to an office chair than a lawn chair; Maxfield, Stalder, & Zelinsky, 2014), and (5) that guidance becomes weaker as targets climb the category hierarchy (the guidance to "race car" is greater than the

guidance to "car," which is greater than the guidance to "vehicle"; Maxfield & Zelinsky, 2012). It is this latter effect of category hierarchy on attention control that was the manipulation of interest in the present study.

Methods

Behavioural data collection

Behavioural data were obtained from Yu et al. (2016) and were collected using the SBU-68 dataset. This dataset consisted of crossly-cropped images of 68 object categories that were distributed across three levels of a category hierarchy. There were 48 subordinate-level categories, which were grouped into 16 basic-level categories, which were grouped into 4 superordinate-level categories. A categorical search task was used, and the participants were 26 Stony Brook University undergraduates. On each trial a text cue designating the target category was displayed for 2500 ms, followed by a 500 ms central fixation cross and then a six-item search display consisting of objects positioned on a circle surrounding starting fixation. Distractors were from random non-target categories and on target-present trials the target was selected from one of the 48 subordinate-level categories. Participants responded "present" or "absent" as quickly as possible while maintaining accuracy, and there were 144 target-present and 144 target-absent trials presented in random order. For each target-present trial, a participant's goal-directed attention guidance was measured as the time taken to first fixate the cued target. Refer to Yu et al. (2016) for full details of the behavioural stimuli and procedure.

Category-consistent features

Previous work used a generative model to predict the strength of categorical search guidance across the subordinate (e.g., taxi), basic (e.g., car), and superordinate (e.g., vehicle) levels of a category hierarchy (Yu et al., 2016). Briefly, its pipeline was as follows. SIFT (Lowe, 2004) and colour histogram features were extracted from 100 image exemplars of 48 object categories, and the Bag-of-Words (BoW; Csurka, Dance, Fan, Willamowski, & Bray, 2004) method was used to put these features into a common feature space. The features most visually representative of each of these categories were then selected, what we termed to be their *Category-Consistent Features* (CCFs). Specifically, responses were obtained for each BoW feature to all the images of each of a category's exemplars, and these responses were averaged over the exemplars and then divided by the standard deviation in the responses to obtain a feature-specific Signal-to-Noise Ratio (SNR). A feature having a high SNR would therefore be one that occurred both frequently and consistently across a category's exemplars. CCFs for each of the categories were obtained by clustering the features' SNRs and selecting the highest.

This BoW-CCF model was able to predict how behavioural performance was affected by target specification at the three levels of the category hierarchy. For example, one specific finding was that the time it took gaze to first land on the target (time-to-target) increased with movement up the hierarchy, what was termed the "subordinate-level advantage." BoW-CCF modelled almost perfectly the observed subordinate-level advantage as a simple count of the number of CCFs extracted for object categories at each hierarchical level; more CCFs were selected for categories at the subordinate level than either the basic or superordinate levels. This result was interpreted as evidence that attention control improves with the number of CCFs used to represent a target category (Yu et al., 2016, should be consulted for more details). The present method adopts the SNR definition of CCFs from Yu et al. (2016), but critically uses VsNet to extract these features (see next section). Also borrowed from the previous work is the method of predicting search guidance from the number of extracted CCFs, a measure that we find desirable in that it is relatively simple and intuitive (more CCFs = better attention control).

Extracting CNN-CCFs

The CCF method selects representative features (which may or may not be discriminative) that appear both frequently and consistently across the exemplars of an object category, but the method itself is largely feature independent. In previous work (Yu et al., 2016) these CCFs were selected from a large pool of BoW features; in our current adaptation we select CCFs from the even larger pool of features from a trained CNN, where each trained convolutional filter is considered a feature and a potential CCF. We

hypothesize that the more powerful CNN-CCF features will represent more meaningful visual dimensions of an object category. For example, whereas BoW-CCFs might have coded the fact that many taxis are yellow and represented the various intensity gradients associated with their shape, a CNN-CCF representation of taxis might additionally capture tires, headlights, and the signs typically mounted to their roofs. We further hypothesize that these richer feature representations, to the extent that they are psychologically meaningful, will enable better predictions of attention control.

The specific CNN-CCF selection process is illustrated in Figure 1 for the taxi category and a hypothetical network. Given an object category with n exemplars of size $m \times m$, and a trained CNN with L convolutional layers each containing K filters, we forward pass all exemplars through the network to obtain an activation profile of size $m \times m \times n$ for every convolutional filter, $Y_k^{(l)}$, where l and k are indices to the layer and filter number, respectively. To remove border artefacts introduced by input padding, the outer 15% of each $m \times m$ activation map is set to zero. Each $Y_k^{(l)}$ is then reduced to a $1 \times n$ vector, $y_k^{(l)}$, by performing global sum-pooling over each image's $m \times m$ activation map. This pooling yields the overall activation of each filter in response to an exemplar image. Having these exemplar-specific filter responses, we then borrow from the BoW-CCF pipeline and compute a SNR for each filter:

$$\mathrm{SNR}_k^{(l)} = \frac{\mathrm{mean}\,(y_k^{(l)})}{\mathrm{std}\,(y_k^{(l)})}, \tag{1}$$

where the mean and standard deviation are computed over the exemplars. Applying this equation to the activation profile from each filter produces a distribution of SNRs. Higher SNRs would indicate stronger and more consistent filter responses, making these filters good candidates for being CCFs. To identify these CCFs we fit a two-component Gamma-Mixture-Model to the SNR distribution, a method similar to Parametric Graph Partitioning (Yu, Hua, Samaras, & Zelinsky, 2013; Yu, Le, Zelinsky, & Samaras, 2015). We use a Gamma distribution because it has been shown to model spiking neuron activity (Li et al., 2017; Li & Tsien, 2017), and we observed that it describes our CNN SNR distributions very well. The CCFs are then defined as the filters having SNRs higher than the crossover point of the two Gamma components. This pipeline for extracting CNN-CCFs was applied on each convolutional layer independently, as filter activations have different ranges at different layers. Of the 500 training and 50 validation

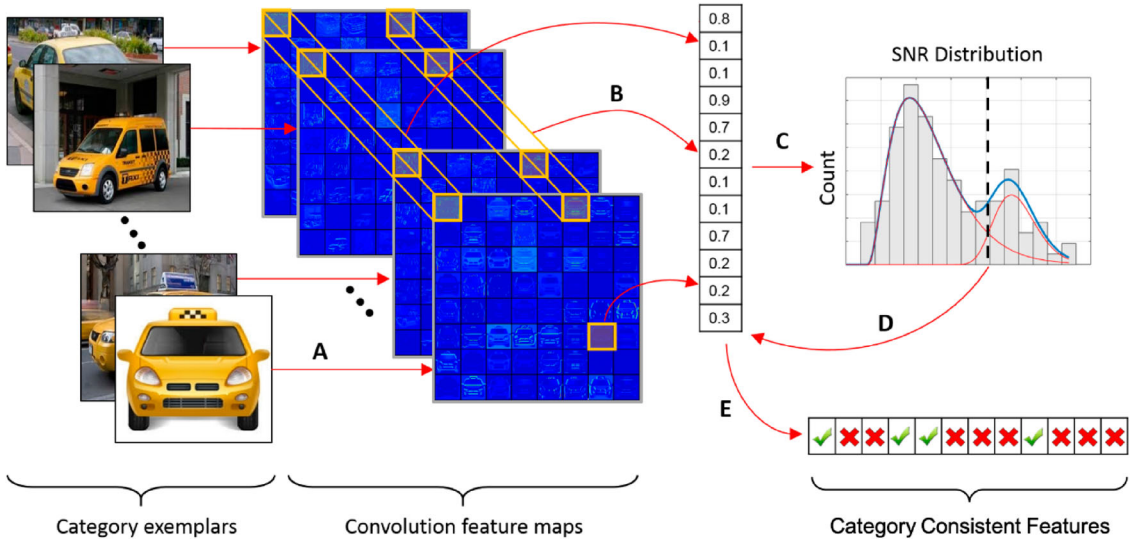


Figure 1. Pipeline of the CNN-CCF extraction method. (A) A set of category exemplars, in this case images of taxis, are input into a trained CNN. (B) Activation maps (or feature maps) in response to each exemplar are obtained for every convolutional filter at each layer. Shown are 64-cell activation maps in a hypothetical layer, where each cell indicates a convolutional filter's response to a given exemplar. In this example, 64 SNRs would be computed (12 shown) by analyzing activation map values for each of the 64 filters across the taxi exemplars. (C) A two-component Gamma mixture model is fit to the distribution of SNRs, (D) and the crossover point determines the CCF selection threshold. (E) Filters having SNRs above this threshold are retained as the CCFs for a given category (√); filters having below-threshold SNRs are dropped (×).

images that were used for each of the 48 tested categories (see the *ImageNet Training* section for details), only the 50 validation images were used to extract a given category's CCFs. The training images were therefore used to learn the filters, whereas the validation images were used to extract the CCFs.

Designing and comparing brain-inspired CNNs

To date, the design of neural network architectures has focused on improving network performance across a range of applications, the vast majority of which are non-biological. Design choices have therefore been largely ad hoc and not informed by either the voluminous work on the organization and function of the primate visual system, or by the equally voluminous literature on visual attention and its role in controlling behaviour. Our broad perspective is that, to the extent one's goal is to understand the primate visual system by building a computational model, it is a good idea to use these literatures to inform the design of new model architectures so as to be more closely aligned with what is known about the primate brain. This is particularly true for the primate visual attention system, where there are rich theoretical foundations in the behavioural and neuroscience literatures that are relatively easy to connect to CNN modelling methods.

VsNet is a rough first attempt to build such a brain-inspired deep neural network, and its detailed pipeline is shown in Figure 2 (top). This effort is "rough" because the neural constraints that we introduce relate only to the gross organization of brain areas along the primate ventral visual stream. There are far more detailed levels of system organization that we could had also considered, but as a first pass we decided to focus on only the gross network architecture. In our opinion this level would likely reveal the greatest benefit of a brain-inspired design, with the expectation that future, more detailed brain-inspired models would only improve prediction of attention control.

Specifically, we designed VsNet to reflect four widely accepted and highly studied properties of the ventral pathway. First, VsNet's five convolutional layers are mapped to the five major ventral brain structures (DiCarlo & Cox, 2007; Kobatake & Tanaka, 1994; Kravitz, Kadharbatcha, Baker, Ungerleider, & Mishkin, 2013; Mishkin, Ungerleider, & Macko, 1983; Serre, Kreiman, et al., 2007). VsNet has a V1, a V2, a combined hV4 and LOC1/2 layer that we refer to as V4-like, a PIT, and a CIT/AIT layer, with these five convolutional layers followed by two fully-connected classification layers. Second, the number of filters in each of VsNet's five convolutional layers are proportional to the number of neurons, estimated by brain surface area (Orban, Zhu, & Vanduffel, 2014; Van Essen et al., 2001), in the corresponding five brain structures. Third, the range of filter sizes at each layer is informed by the range of receptive field sizes for visually responsive neurons in the corresponding structures. And fourth, VsNet differs from other strictly feedforward architectures in that it adopts a brain-inspired implementation of bypass connections based on known connectivity between layers in the primate ventral visual stream. See Figure 2 and the *VsNet Design* section for additional architectural design details.

Our CNN-CCF extraction algorithm is general, and can be applied to the filter responses from any pre-trained CNN. This makes model comparison possible. In addition to extracting CNN-CCFs from VsNet, we used the identical algorithm to extract CNN-CCFs from two other CNNs. One of these was AlexNet (Krizhevsky et al., 2012), a widely used CNN also consisting of five convolutional and two fully-connected layers. Although AlexNet's design was not brain-inspired, it has been used with good success in recent computational neuroscience studies (Cadieu et al., 2014; Hong et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014) and is therefore of potential interest. More fundamentally, it will serve as a baseline against which the more brain-inspired networks can be compared, which is important to gauge broadly how the inclusion of neural constraints in a CNN's design translates into improved prediction performance. We also extracted CNN-CCFs from a model that we are calling Deep-HMAX, our attempt to create a CNN version of the influential HMAX model of object recognition (Serre, Oliva, & Poggio, 2007). HMAX was designed to be a biologically plausible model of how the recognition of visually complex objects might be implemented in ventral brain circuitry (Riesenhuber & Poggio, 1999; Tarr, 1999), but it cannot be fairly compared to more recent and powerful convolutional network architectures. Our Deep-HMAX model keeps the basic architectural design elements of HMAX intact, with the most central among these being the inclusion of simple and

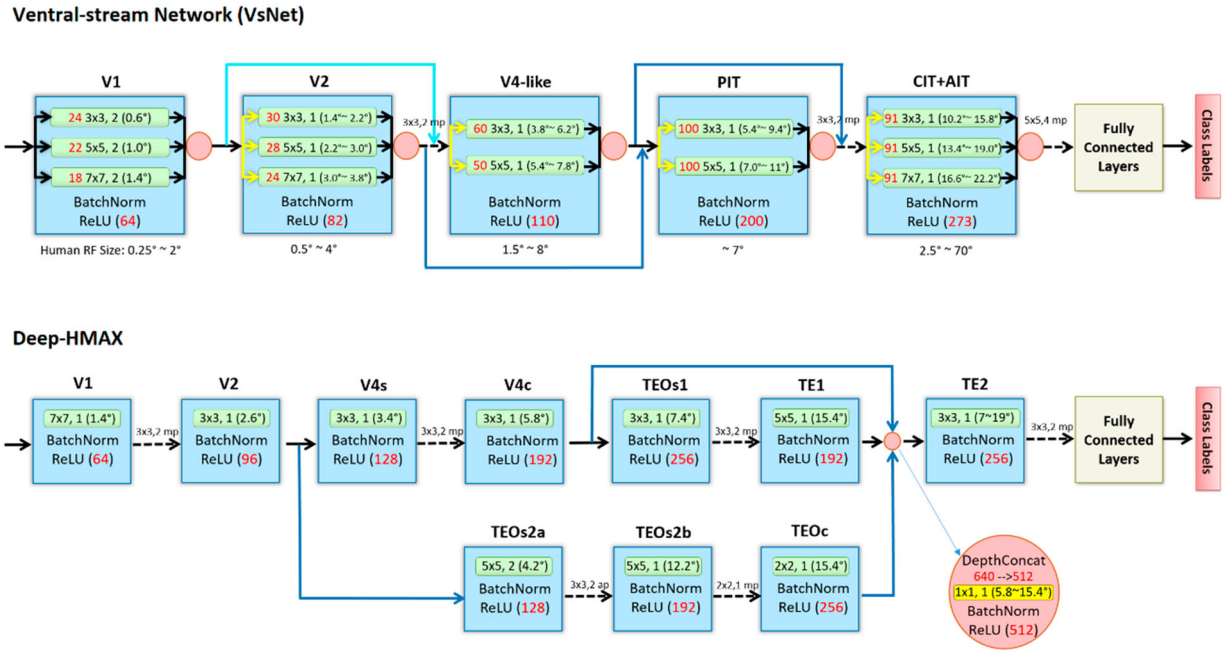


Figure 2. The architectures of VsNet and Deep-HMAX. Each blue box represents a convolutional layer, with the corresponding ventral-pathway area labelled above. Pink circles are Depth-Concat layers that concatenate the input maps from the depth dimension. Arrows indicate input to output direction, dashed arrows represent max-pooling layers and their kernel sizes and strides, yellow arrows represent dimensionality reduction via 1×1 filters, and blue arrows are skip connections that can be either a direct copy (dark blue) or a dimensionality-reduced copy (light blue) via 1×1 filters. Green rectangles within each layer represent a set of filters, where the number of filters is in red, followed by the filter size, stride size, and the corresponding receptive field (RF) size in visual angle shown in parentheses (assuming 1° spans 5 pixels). Note that both VsNet and Deep-HMAX attempt to match the RF sizes of the convolutional filters in each layer to the range of the RF size estimates in each of the five human ventral visual pathway areas. These target RF size ranges are indicated at the bottom of each VsNet layer (see the *Receptive Field Size* section for details on how these estimates were obtained). Each convolutional filter is followed by a Batch Normalization layer (BatchNorm; Ioffe & Szegedy, 2015) and a Rectified Linear activation layer (ReLU).

complex cell units, but replaces the originally hand-crafted units with convolutional layers that learn the simple and complex cell responses from visual input, thereby making possible a more direct comparison to VsNet. Figure 2 (bottom) shows the architecture of Deep-HMAX, and additional details can be found in the *ImageNet Training* section. Broadly, the model has a very different architecture than VsNet, with one example being that it uses 10 convolutional and two fully-connected layers. By comparing Deep-HMAX and VsNet it is therefore possible to see how a fairly gross level of brain organization might affect network performance. Note also that VsNet was computationally disadvantaged in these comparisons because it used the smallest number of convolutional filters to predict attention control; AlexNet has 1152 filters, Deep-HMAX 1760, but VsNet only 726 (excluding 1×1 dimensionality-reduction filters). This conservative design means that, to the extent that VsNet better predicts attention control than the other models, this benefit would likely be due to its brain-inspired architecture and not simply greater computational power.[2]

Vsnet design

VsNet is brain-inspired in three key respects: the number of filters at each convolutional layer is proportional to the estimated number of neurons in the corresponding brain structure, the sizes of filters at each layer are proportional to neuron receptive field sizes in corresponding structures, and the gross connectivity between its layers is informed by connectivity between structures in the primate ventral visual stream. Each of these brain-inspired constraints will be discussed in more detail. With respect to VsNet's broad mapping of convolutional layers to brain structures, the mappings of its first layer to V1 and its second layer to V2 are relatively noncontroversial. However, we wanted VsNet's third convolutional layer to map to V4, a macaque brain area, and identifying a homolog to V4 in humans is less

straightforward. A structure has been identified as "human V4" (hv4), and neurons in this structure are organized retinotopically (Brewer, Liu, Wade, & Wandell, 2005; Fize et al., 2003; McKeefry & Zeki, 1997) like macaque V4, but their feature selectivities are somewhat different. Macaque V4 neurons are selective to colour, shape, and boundary conformation (Cadieu et al., 2007; Desimone, Schein, Moran, & Ungerleider, 1985; Pasupathy & Connor, 2002), whereas neurons in hV4 respond mainly to just colour and occupy a proportionally much smaller cortical surface area (Brewer et al., 2005; Larsson & Heeger, 2006; McKeefry & Zeki, 1997). For humans, shape and boundary and other object-related processing likely occurs in lateral occipital areas 1 and 2 (LO1/2; Larsson & Heeger, 2006). LO1/2 is also retinotopically organized and is anatomically adjacent to hV4 (Van Essen et al., 2001). In an effort to obtain a sufficiently large number of learnable mid-level features, we therefore map VsNet's third convolutional layer to a combination of hV4 and LO1/2, referred to here as "V4-like." Our intent was to map VsNet's deeper layers to IT, and decisions had to be made about these mappings as well. To keep congruence with the monkey neurophysiology literature, we specifically wanted to identify human homologs to macaque TEO and TE. For VsNet's fourth layer we settled on a structure anterior to hV4, termed "human TEO" in (Beck, Pinsk, & Kastner, 2005; Kastner et al., 2001; Kastner, Weerd, Desimone, & Ungerleider, 1998) and PIT elsewhere (Orban et al., 2014), and for its fifth layer we chose central and anterior inferotemporal cortex (CIT + AIT; Rajimehr, Young, & Tootell, 2009), roughly macaque TE. We will show that CNN-CCFs extracted from VsNet, a CNN having this more primate-like architecture, better predicts primate behaviour.

Ventral stream surface areas

The numbers of convolutional filters in VsNet's layers were based on estimates of human brain surface areas in the mapped structures. Specifically, V1, V2 and V4-like surface areas were estimated to be 2323 mm^2, 2102 mm^2, and 2322 mm^2, respectively (Larsson & Heeger, 2006). For PIT and CIT + AIT, we estimated their surface areas to be approximately 9 times larger than the surface areas in the corresponding macaque structures (TEO and TE, respectively; Orban et al., 2014), based on reported differences in cortical size between macaque and human (Van Essen et al., 2001). This resulted in an estimate of PIT having a surface area of 3510 mm^2, and of CIT + AIT having a surface area of 3420 mm^2. Having these surface area estimates, one approach might make proportional allocations of convolutional filters at each layer, but this would ignore the fact that some of these structures have a retinotopic organization. Retinotopy requires that the RFs of neurons having similar feature selectivities are tiled across the visual field in order to obtain location-specific information, and this duplication of neurons is a major factor determining the surface area of some brain structures. CNNs have no retinotopy; their filters are convolved with a visual input rather than duplicated and tiled over an image. To equate the two, we derive a duplication factor that estimates the latent number of uniquely selective neurons within each brain structure, and then makes the number of convolutional filters in the corresponding layer proportional to this estimate. In doing this we make a simple assumption. If the average RF size for a neuron in a ventral stream structure is as large as the entire visual field, then there would be no need for the retinotopic duplication of this type of neuron for the purpose of capturing information from across the visual field. This would lead to a duplication factor of 1. However, if in this example the average RF size for a neuron covers only a quarter of the visual field, then there would minimally need to be four neurons of this type organized retinotopically to cover the entire visual field. This would lead to a duplication factor of 4. More generally, the following formulas were used to calculate the duplication factor for a given ventral stream structure and to determine the number of convolutional filters in VsNet's corresponding layer:

$$\#\,\text{filters} \propto \frac{\text{surface area}}{\text{duplication}},\quad \text{duplication} = \log\left(\frac{\text{visual area}}{\text{RF size}}\right), \tag{2}$$

where both the area of the visual field and neuron RF size are expressed in degrees squared. We take the log of these values' proportion in order to scale down the increase in the numbers of filters from lower to higher layers, so as to stay within hardware constraints. For the current implementation, 1° of visual angle equalled 5 pixels, making the 224×224 pixel input images subtend approximately $45° \times 45°$ of visual

area in the model's "field of view." For each ventral stream area, we then take the average RF size at 5.5° eccentricity to be representative of neuron RF sizes in that structure (i.e., we currently do not capture the foveal-to-peripheral increase in RF sizes within a structure, due to computational limitations, but see the *Receptive Field Size* section below). Doing these calculations, we obtained the representative RF size estimates of 1°, 3°, 5°, 7°, and 12° for V1, V2, V4-like, PIT, and CIT + AIT, respectively (see also Kravitz et al., 2013). Finally, using these values in the duplication factor calculation, and setting the total number of filters in the first convolutional layer (V1) to 64 (to be directly comparable to AlexNet), we obtain the final VsNet architecture consisting of 64, 82, 110, 198, and 272 filters across its 5 convolutional layers, excluding 1×1 dimensionality-reduction filters (see Figure 2, top).

Receptive field size

In primates, the RFs of visually-responsive neurons increase in size with distance along the ventral stream; neurons in structures early in this pathway have small RFs, whereas those in later structures have larger RFs (Freeman & Simoncelli, 2011; Kravitz et al., 2013). Moreover, within visual structures preserving reasonable retinotopy (V1 to V4) cortical magnification causes neurons coding the central visual field to have relatively small RFs, and neurons coding increasingly peripheral locations to have increasingly larger RFs (Engel, Glover, & Wandell, 1997; Freeman & Simoncelli, 2011; Harvey & Dumoulin, 2011; Wade, Brewer, Rieger, & Wandell, 2002). VsNet was designed to grossly capture both of these properties. However, this latter relationship between RF size and visual eccentricity is difficult to implement in a CNN, where models are computationally constrained to have filters of only a single size within each of their convolutional layers (He et al., 2016; Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; Zeiler & Fergus, 2014), with a current exception being the Inception Module from Szegedy et al. (2015). This is because the convolutional filters in a CNN were specifically designed *not* to be applied at specific image locations (i.e., shared weights), making the modelling of a changing retinotopy difficult. But we approximate the variability in RF sizes due to scaling with eccentricity, and we do this by using parallel sets of 3×3, 5×5, and 7×7 pixel convolutional filters in each of VsNet's layers (except for layers 3 and 4, which used only 3×3 and 5×5 filters). These sizes were chosen so as to approximate the range of RF sizes within each of the corresponding structures (Harvey & Dumoulin, 2011; Kastner et al., 2001; Rousselet, Thorpe, & Fabre-Thorpe, 2004; Smith, Williams, & Greenlee, 2001). Given our use of 5 screen pixels to represent 1° of visual angle, the 224×224 pixel ImageNet images used for training subtended a visual angle of 45°. More importantly, a 3×3 filter In VsNet's V1 layer spanned 0.6°, a 5×5 filter spanned 1°, and a 7×7 filter spanned 1.4°. This range of RF sizes (0.6° to 1.4°) maps closely onto the range of RF sizes in V1 (0.25° to about 2°). These filters are convolved with the input, producing feature maps that we concatenate in depth, such that the convolutional filters at the next higher layer (V2) receives responses from filters having three different sizes. For example, the stacking of layer 2's 3×3 filters on top of layer 1's 3×3, 5×5, and 7×7 filters, results in layer 2's 3×3 filters having RF sizes of 1.4°, 1.8°, and 2.2°, respectively (the parenthetical values listed in Figure 2 for VsNet's 3×3 V2 filters). Doing this also for the 5×5 and 7×7 filters produced a range of sizes again corresponding well to the range of RF sizes observed in V2 neurons (the values below each blue box in Figure 2, top). A similar procedure was followed for VsNet's V4-like layer, which produced similarly good estimates of neuron RF sizes. Over VsNet's first three layers, the filters at each higher layer therefore had, not only larger RFs, but also a broader range of RF sizes. For VsNet's PIT and CIT + AIT layers, the same numbers of filters were allocated in the parallel sets, reflecting the relaxation of a retinotopic organization in the corresponding ventral structures. Note also that GoogLeNet's Inception Module (Szegedy et al., 2015) has a similar parallel-filter architecture, but it is unlikely that the design of this model was inspired by the primate visual system.

Bypass connections

In addition to the feed-forward projections that connect each lower-level ventral stream area with areas at the next higher level along the pathway, good evidence also exists for connections that skip or bypass neighbouring ventral structures (Kravitz et al., 2013; Nakamura, Gattass, Desimone, & Ungerleider, 1993; Nassi & Callaway, 2009). VsNet captures both types of ventral stream connectivity, although it should be considered only a first-pass attempt to

do so; capturing the minutia of this brain connectivity is currently beyond its scope. The direct connections are already embedded in its feed-forward design, so the focus here will be on detailing its bypass connections. Major bypass connections exist from V2 to TEO (Nakamura et al., 1993; Tanaka, 1997) and from V4 to TE (Tanaka, 1997), with a weaker bypass connection known to exist between V1's foveal region to V4 (Gattas, Sousa, Mishkin, & Ungerleider, 1997; Nakamura et al., 1993; Ungerleider, Galkin, Desimone, & Gattass, 2007). These three bypass connections were designed into VsNet. We added a weak bypass connection from layer 1 (V1) to layer 3 (V4-like), a full bypass from layer 2 (V2) to layer 4 (PIT), and another full bypass from layer 3 (V4-like) to layer 5 (CIT + AIT). We implemented these bypass connections by concatenating in the depth dimension the lower layer's output to the target layer's input. Note that this concatenation method is different from the summation method used by ResNet (He et al., 2016), but is conceptually similar to the Inception Module design used by GoogLeNet (Szegedy et al., 2015). Following Szegedy et al. (2015), we also use 1×1 filters before each of VsNet's convolutional layers (except layer 1, where they are not needed) for dimensionality reduction and memory conservation (yellow arrows in Figure 2, top). We chose this concatenation method in order to give VsNet maximum flexibility in how bypassed information is best combined with information at the target layer, which we believe is preferable to assuming that the cortex simply sums this information. Specifically, a full bypass was implemented by concatenating in the depth dimension a complete copy of the source layer's output feature map to the end of the target layer's input map. We implemented a weak bypass similarly, but now the source layer's output map was depth-reduced (dimensionality reduced by half via 1×1 convolutional filters) before being concatenated with the target layer's input feature map.

Imagenet training

VsNet, AlexNet, and Deep-HMAX were trained using ImageNet (Russakovsky et al., 2015). All training and validation images were resized to have the shortest side be 256 pixels while keeping the original aspect ratio, and the standard data augmentation methods of random crops (224×224) and random horizontal flips were employed. Centre crops were used to compute validation accuracies at the end of each training epoch. The training batch-size for AlexNet, Deep-HMAX, and VsNet was 128, 64, and 60, respectively, determined by GPU limitation. Each network was trained using 4-threads, with image data stored on a solid-state drive, and 60 training epochs took roughly 2–4 days to complete using a 2.93 Ghz Intel Xeon x3470 processor with 32 Gb of memory and a single Titan X GPU. Networks were implemented using Torch7, and the method from He, Zhang, Ren, and Sun (2015) was used for parameter initializations.

Following ImageNet training, networks were fine-tuned using the SBU-68E dataset (https://github.com/cxy7452/CNN-CCF/tree/master/SBE-68E/), an expanded version of the SBU-68 dataset. The original SBU-68 dataset contained 4800 images of objects, which were grouped into 100 exemplars from each of 48 subordinate-level categories (Yu et al., 2016). These images were further combined hierarchically to create an additional 16 basic-level categories and 4 superordinate-level categories, yielding 68 categories in total. The expanded SBU-68E dataset built on the earlier dataset by exploiting Google, Yahoo, and Bing image searches to obtain 1500 exemplars from each of the same 48 subordinate-level categories, thereby making it more suitable for deep network training. GIST descriptors (Oliva & Torralba, 2001) were used to meticulously remove image duplicates, followed by a manual pruning of the images to ensure that those with incorrect class labels were removed and that the retained images were well-cropped around the labelled object. These exclusion criteria yielded 500 training and 50 validation images per category, for a total of 24,000 training and 2400 validation images in the expanded set. All images were resized such that the shortest side was 256 pixels wide while retaining the original aspect ratio.

Results

CNN-CCFs predict visual attention control

Although all of the implemented models were trained for object classification and therefore have no dimension of time, we exploited a finding from Yu et al. (2016) that allowed us to relate model performance to search efficiency. This study showed that the number of BOW-CCFs extracted from their model accurately predicted

the time that participants took to first fixate a target category cued at each of the three tested hierarchical levels. Our first evaluation was therefore to obtain CNN-CCFs from the convolutional layers of identically-trained VsNet, AlexNet, and Deep-HMAX networks using the previously described CNN-CCF feature selection method, and determine whether the number of CNN-CCFs extracted from these three deep networks also predicted the effect of a target category's hierarchical level on attention control. As shown in Figure 3(A), all of the models tested were highly successful in capturing the behavioural trend of increasing time-to-target with movement up the category hierarchy. This demonstration is important in showing that the number of CCFs is highly generalizable in its ability to predict the effect of hierarchical level on categorical guidance; across four very different models, more CCFs were extracted for subordinate-level categories compared to basic, and for basic-level categories compared to superordinate, with the assumption that greater numbers of these features form better target templates that can more efficiently guide attention to the target categories.

Capturing the behavioural trend in guidance across category levels is one thing, using CCFs to predict attention control to individual categories is a different and far more challenging goal. Our experimental logic, however, is the same; the more CCFs that can be extracted for a target category, the better attention should be able to bias the features representing an unseen exemplar of that category in the visual input, with the behavioural expression of this biasing being the guidance of gaze to the target's location. Across categories we therefore predict a negative correlation between the number of CCFs and the search time-to-target measure, with more CCFs leading to shorter target fixation times. But recall that each network layer is extracting its own CCFs, and it is unreasonable to believe that the attention control mechanism would disregard network depth and weigh all of these features equally. We therefore found a CCF weighting across each network's convolutional layers that optimized a correlation (Spearman's ρ) between the number of CCFs extracted at each layer and the time-to-target measure of attention control, with each network model having its own optimized layer weights. The advantage of this formulation is that it allowed Spearman's ρ to be used directly as an objective function to optimize the layer-wise weights, W, which we did using beam search with random steps (Vicente, Hoai, & Samaras, 2015).

Figure 3(B) shows these category-specific predictions of guidance efficiency at each hierarchical level and for the four tested CCF models. Note that prediction success is indicated by higher negative correlations, plotted upward on the y-axis. Predictions from the BoW-CCF model were poor for subordinate

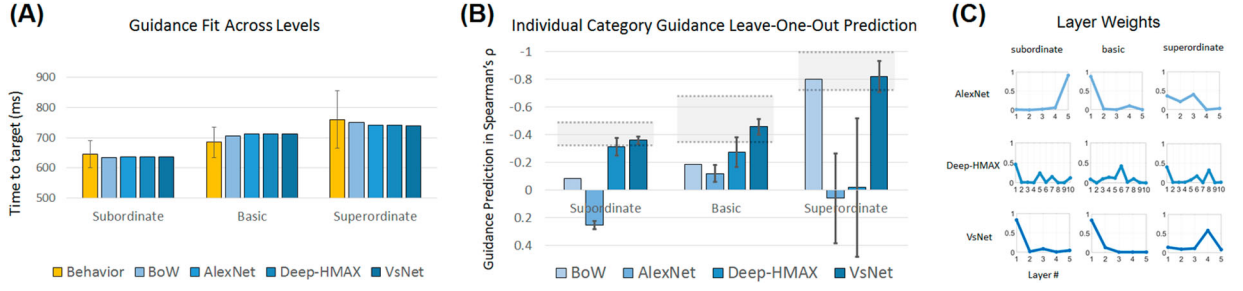


Figure 3. (A) Predictions of human attention control (time to fixate the target) for one CCF model using Bag-of-Words (BoW) and three CNN-CCF models: AlexNet, Deep-HMAX, and VsNet. Results are grouped by level in the categorical hierarchy, and model performances are linearly scaled to best fit the behaviour (i.e., the models' results are put in the behavioural scale). All four models successfully predicted the subordinate-level advantage in attention guidance to categorical targets. (B) Model predictions of attention control to individual target categories within each hierarchical level, evaluated using the leave-one-out method. Given the inverse correlation between number of CCFs and the time needed to fixate a target (Yu et al., 2016), more negative correlations indicate better predictions of attention control. Grey regions indicate performance ceilings on how well a model can predict attention guidance, based on +/− one standard deviation in mean guidance from a "subject model." The subject model was also computed using leave-one-out, only now we found the Spearman's ρ between $n-1$ participants and the participant who was left out (the mean and standard deviation was obtained by repeating this for all participants). The results show that models using BoW or AlexNet to extract CCFs are unable to predict human behaviour. However, the more brain-inspired CNN-CCF models perform better, with VsNet being the best and on par with a human subject model. (C) Best-fitted weights by convolutional layer for each CNN-CCF model, grouped by hierarchical level. VsNet's weight distribution suggests that categorical guidance at both subordinate and basic levels is driven by low-level features, while guidance at the superordinate level is driven by high-level features.

and basic-level categories and significantly worse than those from VsNet and Deep-HMAX. A very good prediction was obtained at the superordinate level, but given that there were only four categories at this level a high correlation might simply have resulted from chance. Interestingly, the number of CNN-CCFs extracted from the widely-used AlexNet model failed entirely in predicting attention guidance to individual target categories. Of the two evaluated brain-inspired CNNs: prediction success from Deep-HMAX was not reliably different from VsNet at the subordinate level ($p = 0.059$), was significantly lower than VsNet at the basic level ($p < 0.001$), and non-existent at the superordinate level, while VsNet's predictions remained very good. Indeed, for individual categories at all three hierarchical levels, VsNet's predictions were well within the performance ceilings (grey regions) computed by having $n-1$ participants predict the behaviour of the participant left out. This means that VsNet's predictions were as good as can be expected given variability in the participant behaviour, and it is the only model of the four tested for which this was consistently the case. These results suggest that not all brain-inspired CNNs are created equal; a CNN designed after the ventral visual pathway is more predictive of attention control than the architecture used in Deep-HMAX.

CNNs have been criticized as being "black boxes"; they perform well but the reason for their success defies understanding. We prefer to think of CCNs as "transparent boxes," ones that can be probed and peered into in attempts to decipher how they work.[3] As one example, Figure 3(C) plots the optimized layer weights (W), grouped by level in the category hierarchy, for each of the three CNN models tested. For subordinate and basic-level targets, VsNet's distribution of CCF weights showed a very clear dominance for early visual features, meaning that these features best predicted attention control for targets specified at these categorical levels. However, for superordinate-category targets the CCFs learned by its CIT + AIT layer were the most predictive. VsNet's CCF layer weighting therefore captures what has become a core finding in the categorical search literature; that lower-level features dominate attention control when relatively clear visual properties of the target can be discriminated (as is often likely to be true for subordinate and basic-level targets), but that higher-level features must be used to control attention to targets that do not have clearly representative visual properties (such as targets specified at the superordinate level). In contrast, the optimized CCF layer weightings for Deep-HMAX across its 10 layers seemed more erratic (although perhaps suggesting the emergence of a pattern similar to VsNet), and the very weak correlations from AlexNet made its optimized CCF layer weights uninterpretable. Once again, the type of brain-inspired design of the CNN matters. Here we show that differences in CNN architectures have consequences, one of which is that different CCFs are extracted at different layers. We believe that the distribution of CCF layer weights in the CNN "box" is a meaningful pattern that we can observe, analyse, and potentially use to generate predictions for further behavioural studies of attention control.

CNN-CCF visualization

VsNet is also a transparent box in that it is possible to peer inside to see what image patterns its CCFs were coding – the representative visual features of an object category. The CCFs for a given category can be visualized by finding the regions in input images that best activate a given CCF (a particular convolutional filter). Specifically, we first forward-pass an image of a category exemplar through VsNet to obtain the maximally-responsive locations in a feature map for the CCF of interest, and then probe backwards from the filter's most activated location to the pixels in the image that was causing this maximal response (Zeiler & Fergus, 2014). Figure 4(A) visualizes the image regions eliciting the five largest responses from CCFs, based on a ranking of their SNR scores, at each of VsNet's layers for one exemplar image from the taxi category. Note that these maximally-active CCFs seem in some cases to be representing object parts that are specific to typical taxis, such as the rooftop advertisements, but also parts that are more broadly representative of cars, such as wheels, windows, and side mirrors. This observation also illustrates the generative nature of CCFs; they code the features that are common to a category (rooftop signs and wheels, in the case of taxis) regardless of whether these features are discriminative (police and race cars also have wheels, but only taxis have rooftop signs).

The aggregated locations of maximally-active CNN-CCFs can also be used to detect categories of objects.

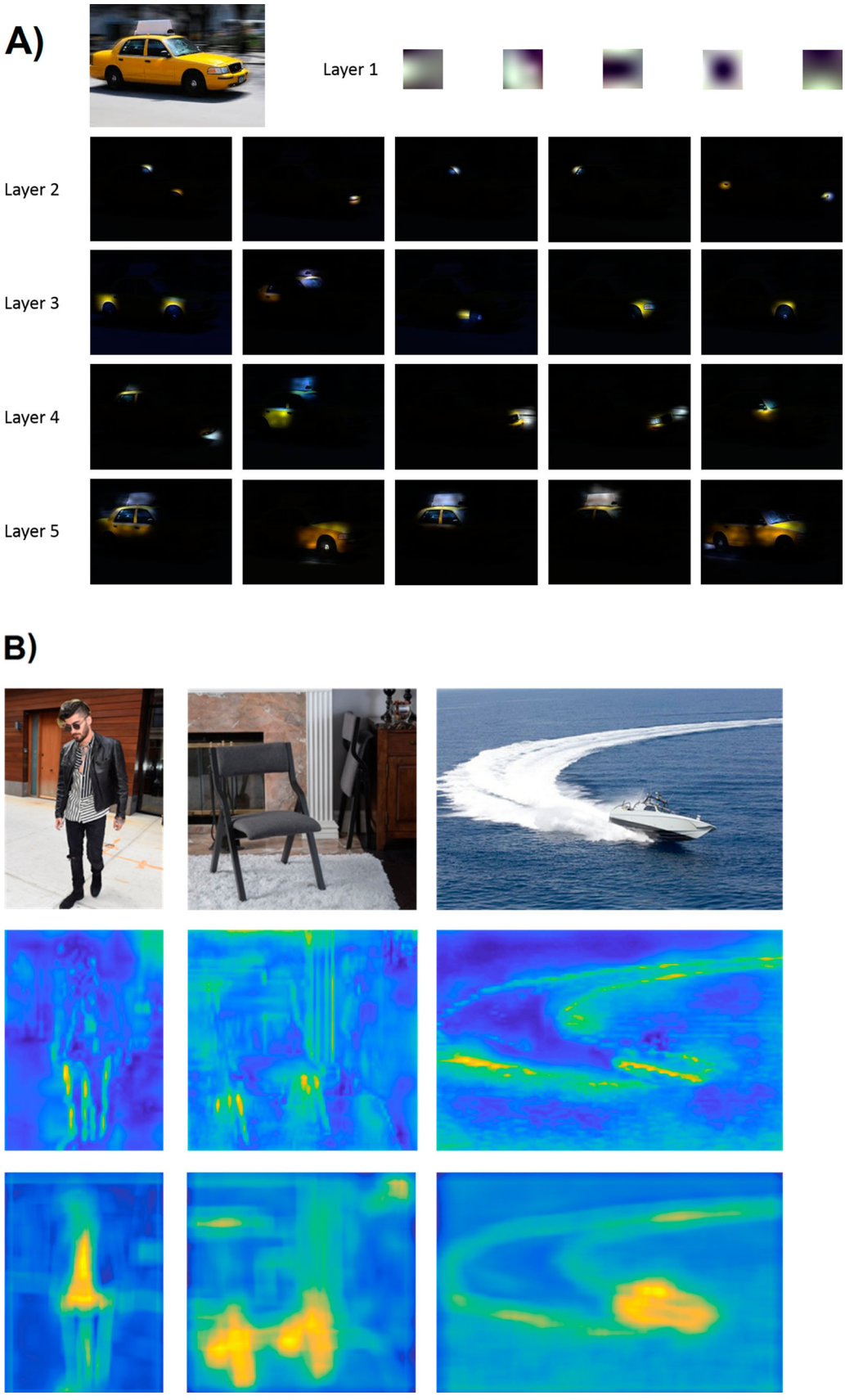


Figure 4. (A) A CNN-CCF visualization for the taxi category. The visualized patches are the top 5 CCFs, ranked based on their SNR, for each convolutional layer in VsNet in response to one taxi image exemplar (upper-left). The CNN-CCFs seem to preferentially code object parts that are representative of a typical taxi, such as tires, headlights, windows, and the rooftop sign. (B) Examples of CNN-CCF activation maps used for object detection: original images (top row), activation maps from AlexNet (middle row), and activation maps from VsNet (bottom row). Activation maps show the combined activity for a given category's CNN-CCFs, where warmer colours indicate greater activation. Categories from left to right: shirt (basic), folding chair (subordinate), and speedboat (subordinate).

This is because CCFs broadly capture the different parts of an object category, at different scales, making it possible to detect the presence of a target object category in an image simply by detecting its constituent representative parts. As qualitative examples, Figure 4(B) shows images depicting the object categories of shirt, folding chair, and speedboat (top row), paired with the combined activation maps from CCFs extracted for those categories by VsNet (bottom row). None of these images were part of the training set. Note that the CCFs for the shirt category precisely differentiate that object from the categorical siblings of jacket and pants, and that the CCFs for the category of folding chair are clearly coding the chair's legs, which is a part that discriminates that subordinate-level category from other chairs. The speedboat example illustrates the difference between bottom-up saliency and top-down goal-directed attention control; CCFs activate strongly to the small boat but almost not at all to its far more salient white wake. Also significant about this demonstration is that this precise object localization was accomplished simply by combining the CCF activation maps without the need for additional processing costs. Object detection was, in a sense, free (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2014).

For comparison, we also plot in Figure 4(B) localizations using CNN-CCFs extracted from AlexNet (middle row). Note that we did not do this for Deep-HMAX due to the more complex architecture of that model making this localization method difficult to implement. The comparable CCFs visualized from AlexNet were more hit or miss; sometimes they were reasonable (e.g., folding chair) whereas other times the visualizations were clearly subpar (e.g., speedboat) or inexplicable (e.g., shirt). This suggests that not all CCN-CCFs are created equal, with those from VsNet producing consistently better object detection across these examples than those from AlexNet, speculatively due to VsNet's brain-inspired design. However, we believe that the broader message from this analysis is that CCFs look generally good when visualized, and although VsNet may be more successful in this regard than AlexNet they both produced reasonable results because they both extracted CNN-CCFs.

Large-scale image classification

Prior to extracting CCFs from the three CNN models, the networks must be trained to learn an initial set of features. This initial training, and later validation, was done using ImageNet (Russakovsky et al., 2015) following standard training procedures (see the *ImageNet Training* section for details). Although not directly within this study's question of focus (goal-directed attention control), comparing classification performance from initial training results can indicate feature quality differences between the different network architectures. Given their ready availability, we therefore discuss these results briefly here. Another motivation for raising this topic is that we believe VsNet produced interesting behaviour, and this might make it also of interest to readers doing large-scale object classification.

Table 1 summarizes model performance on ImageNet's validation set. Classification accuracies for all models were high, indicating that the networks were successfully trained. However, VsNet achieved the highest classification accuracy while AlexNet achieved the lowest. This is a notable finding because VsNet's design was specifically engineered after the ventral stream of visual brain areas, and was specifically *not* designed or optimized for classification accuracy. Yet, it outperformed a model that *was* designed to optimize classification – AlexNet – by what is considered a significant margin in the computer vision literature. Moreover, although deeper CNN architectures generally outperform shallower networks (Simonyan & Zisserman, 2015), in VsNet we found an exception to this rule. While Deep-HMAX, the deepest network of the three, outperformed AlexNet in classification accuracy, it was less accurate than the shallower VsNet.

Also notable is the fact that VsNet achieved the highest accuracies despite having fewer convolutional filters than either Deep-HMAX or AlexNet (726, 1760, 1152 filters, respectively, excluding 1×1 dimensionality-reduction filters). Similar to network depth, the number of non-1×1 convolutional filters typically correlates highly with network performance in the computer vision literature: ResNet (He et al., 2016) has

Table 1. Top-1 and top-5 validation accuracies for the three CNN models on the ImageNet dataset (Russakovsky et al., 2015). The overall high accuracies indicate the low likelihood of overfitting. Note that network performance improved with the degree of brain inspiration in its design.

ImageNet	AlexNet	Deep-HMAX	VsNet
Top-1 Accuracy	57.7%	59.6%	61.5%
Top-5 Accuracy	80.6%	82.4%	83.9%

more filters and is more powerful than GoogLeNet (Szegedy et al., 2015), which is larger and more powerful than VGG (Simonyan & Zisserman, 2015), which is larger and more powerful than AlexNet (Krizhevsky et al., 2012). Although VsNet is not alone in this regard (see Canziani, Culurciello, & Paszke, 2017), it is an example of a CNN outperforming networks having many more convolutional filters. This reversal of trend suggests that VsNet was able to learn better representations despite having a smaller pool of features from which to sample. We believe that this greater convolutional kernel efficiency (Figure 5) is a meaningful benefit of VsNet's brain-inspired design.

Discussion

This study is among the first to use a CNN to predict a goal-directed, attention-controlled behaviour – the guidance of gaze to a categorically-defined target (see also Adeli & Zelinsky, 2018). Previous work showed that a computationally-explicit generative model (BoW-CCF) could predict a relationship between a target's level in a category hierarchy and the efficiency of attention control (Yu et al., 2016), but the predictive success of this BoW-CCF model was limited to only three hierarchical levels, essentially three data points. When this model attempted to predict the degree of attention guidance to individual target categories, it failed (see Figure 3(B)). However, the BoW method has been largely replaced in recent years by deep learning models, which are able to learn far richer feature representations (Razavian, Azizpour, Sullivan, & Carlsson, 2014) that lead to significantly better performance in large scale image classification (Krizhevsky et al., 2012; Sanchez, Perronnin, Mensink, & Verbeek, 2013). Here we show that *category-consistent features* (CCFs) selected by a CNN can predict, not just the overall effect of hierarchical level on attention control, but also the degree of attention control to individual target categories across this same three-level hierarchy. This generalization across categories is testimony to the robustness of CCFs that are extracted from a CNN model. We consider this CNN-CCF method to be the more important and lasting contribution of this work. Computational models of cognitively complex behaviours, to the extent they are successful, tend to have short lives (e.g., http://saliency.mit.edu/results_mit300.html), and we expect that VsNet will soon be replaced by even more brain-inspired models. However, we believe that the selection of category-consistent features from a deep network, and the use of these features to predict attention control, are ideas that might drive research in attention modelling for years to come.

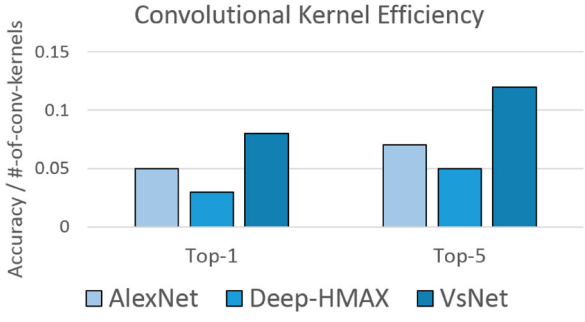


Figure 5. Plotted is a measure of the efficiency of a convolutional kernel (filter), defined as accuracy per convolutional filter. Kernel efficiency is shown for Top-1 (percentage of cases in which the model's most confident prediction was correct) and Top-5 (percentage of cases in which the correct object category was among the five-most confident predictions from the model) evaluation metrics. VsNet was found to have the highest convolutional kernel efficiency, followed by AlexNet. Deep-HMAX was the least efficient network, possibly due to its long parallel branches learning redundant features.

VsNet is (currently) significant in that it is an artificial deep network whose design is broadly informed by the architecture of the primate ventral visual stream. We compared VsNet to another brain-inspired model (Deep-HMAX) and a popular, yet brain-uninspired model (AlexNet), and showed that VsNet best predicted attention control. Computationally, this demonstration is significant in two respects. First, AlexNet has 58% more learnable filters than VsNet, and Deep-HMAX has 142% more, yet VsNet was more predictive than both. Although similar exceptions have been noted (Canziani et al., 2017), the far more common relationship is that prediction success increases with the number of convolutional filters (ResNet > GoogLeNet > VGG > AlexNet). VsNet violates this general trend by predicting attention control better than more powerful networks having significantly more convolutional filters. Second, given that deeper CNN architectures generally outperform shallower architectures (Simonyan & Zisserman, 2015), it is notable that Deep-HMAX, a 10-layer CNN, did not compare more favourably to VsNet, a network with half its depth. We speculate that this might be due to Deep-HMAX's long parallel branches (forming

after its V2 layer) creating a redundancy in the filters that it learns, in contrast to VsNet that uses short bypass connections to route lower-level information more directly to higher layers. Indeed, it is plausible that this more direct routing of visual inputs from lower to higher layers is exactly what was responsible for VsNet's better prediction of attention control. Determining the specific sources of VsNet's success, and why other brain-inspired designs fail, will be an important direction for future work. For now we must limit our conclusion to the fact that VsNet, a comparatively simple network in terms of its number of trainable filters, learned target-category representations that best predicted human attention control in the tested search task. We further speculate that the CNN-CCFs learned by VsNet approximate the local target-feature circuits learned along the ventral visual stream that form the representations that can be biased by a goal-specific attention target. VsNet does not, however, currently address the computational mechanism by which the attention target biases these features, and adding this will be another direction for future work (see also Adeli & Zelinsky, 2018; Zhang et al., 2018).

A theoretically important implication of our study is that different model architectures produced different CCFs at different layers, and that this impacted the prediction of human search behaviour. These CCF differences occurred despite all the models being able to classify the target categories quite well. The suggestion from VsNet is that early and intermediate-level visual features – those preceding the features used for large-scale object classification – can be biased for the purpose of directing attention to the location of a target goal. Note that this restates a view of attention control that is widely accepted in the visual search literature (Wolfe, 1994; Zelinsky, 2008), although there are more recent suggestions that search may additionally be guided by recognized objects (Einhäuser, Spain, & Perona, 2008) and scene context (Neider & Zelinsky, 2006). VsNet, DeepHMAX, and AlexNet could all classify the tested object categories, but only the CNN-CCFs extracted from VsNet predicted the guidance of overt attention to the locations of these categories in visual inputs. Why? The importance of intermediate levels of representation is sometimes lost in the modelling literature, where the goal is often to achieve higher classification accuracy over an increasingly broad range of categories. However, high classification accuracy does not mean that a particular architecture has learned good representations for guiding overt attention. It may be that these higher-level representations become highly unreliable for peripherally-degraded visual inputs (Zelinsky, Peng, Berg, et al., 2013), a speculation that dovetails well with the very rapid decline in human object recognition ability with increasing visual eccentricity (Nelson & Loftus, 1980; Thorpe, Gegenfurtner, Fabre-Thorpe, & Bülthoff, 2001). If objects are difficult to recognize in the degraded visual periphery, this would necessarily lessen the contribution of higher, object-based layers of a network to guide search. It is also possible that VsNet learned rich feature representations that are capable of obtaining evidence for the target category in a peripherally-degraded visual input, and that this translated into its better ability to predict search behaviour. Future work will selectively "lesion" VsNet by relaxing its biological constraints for select layers and observing the effect of this lesion on predicting attention control. Such a lesion study approach will be valuable in experimentally isolating the CNN-CCFs that can be biased by goal-directed attention. For now, our take-away message is simple: if one's goal is to model primate attention control, it makes sense to design your model's architecture to be more like that of the visual system embodied in the primate brain.

Far from being a black box, VsNet's use of CNN-CCFs hints at how attention control and object classification processes might interact across the layers of a deep network. One clue comes from the observation that the CNN-CCFs extracted by VsNet segregated reasonably across its layers, with CCFs for subordinate and basic-level categories extracted at the lower layers and CCFs for superordinates extracted at the higher layers. Another clue comes from the promising object localization made possible by projecting CNN-CCF activation from higher layers back to lower layers. This method exploits the high-resolution spatial information coded by lower-level target features (in the early visual layers) to estimate the spatial inputs used by the rich higher-level representations learned for good classification. The demonstration also has implications for Biased-Competition Theory (Desimone & Duncan, 1995; Tsotsos et al., 1995). It shows how a top-down bias can effectively delineate a target object goal in space, which is a

prerequisite computation for the selective routing of visual inputs from that region to higher brain areas for classification. Under this framework, attention control and classification are therefore locked in a cycle of optimization; better classification leads to stronger attention control, which in turn leads to better classification. We believe that the most important function of attention control is to mediate good object classification, and that an understanding of goal-directed behaviour requires that the currently largely separated attention control and object classification literatures be conceptualized as two parts of a broader object-interaction system (Allport, 1980). Brain-inspired CNN models are a promising tool to begin understanding the neurocomputational architecture needed to interact with real-world objects.

Notes

1. In referring to "attention control" we draw a distinction between an "attention target", which we define as the high-level semantic representation specifying an immediate behavioural or cognitive goal (e.g., a designated target category in a search task), and "target features", which we define to be the lower-level visual features representing the attention target in a perceptual input. It therefore follows that "attention control" is the goal-specific biasing of lower-level target features for the purpose of controlling an interaction with the attention target (e.g., directing the fovea to the location of a target goal in a visual input), and a measure of attention control is one that evaluates the success or efficiency in achieving this goal (e.g., the time required to align the fovea with the target). Understanding both the attention target and the target features is essential to understanding attention control and goal-directed behaviour. There could be a top-down attention target reflecting a desire to find a Pekin duck in a pond, but this visual goal could not be realized unless features of Pekin ducks have been learned by the visual system and are therefore available for top-down biasing.
2. Note that our model quantification treats two filters as equivalent in complexity despite their having different numbers of free parameters. Quantifying model complexity in terms of free parameters is arguably not meaningful with respect to higher-level perceptual and cognitive behaviours, such as categorical search. It essentially places more weight on the number of connections comprising a representation rather than the representation itself. A 100×100 filter covers more area (and has 9900 more parameters) than a 10×10 filter, but the larger filter is not likely to be $10\times$ more predictive than the smaller filter; the information coded by the two simply differs in scale and type. Moreover, applying such a quantification scheme to the visual system would imply that the responses from neurons having large receptive fields are more predictive than the responses of neurons having smaller receptive fields, when the opposite is arguably more likely to be true. Our quantification approach draws an equivalency between filters (the units at each layer) and features (or feature maps, as these are derived by convolution with a filter), which is how complexity has historically been conceptualized in the attention literature.
3. We thank Aude Oliva for this term, which she used in a personal communication at the Vision Sciences Society annual meeting (May, 2017); see also Bau, Zhou, Khosla, Oliva, & Torralba, 2017, for a similar sentiment.

Acknowledgements

Invaluable feedback was provided by the members of the Computer Vision Lab and the Eye Cog Lab at Stony Brook University, and by Dr. Talia Konkle and the members of the Harvard Vision Sciences Lab.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Adeli, H., Vitu, F., & Zelinsky, G. J. (2017). A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *The Journal of Neuroscience*, *37*(6), 1453–1467.

Adeli, H., & Zelinsky, G. (2018). Deep-BCN: Deep networks meet biased competition to create a brain-inspired model of attention control. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRw)* (pp. 1932–1942).

Allport, D. A. (1980). Attention and performance. In G. Claxton (Ed.), *Cognitive psychology* (pp. 112–153). London: Routledge & Kegan Paul.

Ballard, D. H., & Hayhoe, M. M. (2009). Modelling the role of task in the control of gaze. *Visual Cognition*, *17*, 1185–1204.

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of computer vision and pattern recognition (CVPR 2017)*.

Beck, D., Pinsk, M., & Kastner, S. (2005). Symmetry perception in humans and macaques. *Trends in Cognitive Sciences*, *9*, 405–406.

Brewer, A., Liu, J., Wade, A., & Wandell, B. (2005). Visual field maps and stimulus selectivity in human ventral occipital cortex. *Nature Neuroscience*, *8*, 1102–1109.

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2016). *What do different evaluation metrics tell us about saliency models?* arXiv:1604.03605.

Cadieu, C., Hong, H., Yamins, D., Pinto, N., Ardila, D., Solomon, E., … Bethge, M. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*, e1003963.

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C., Riesenhuber, M., & Poggio, T. (2007). A model of V4 shape selectivity and invariance. *Journal of Neurophysiology*, *98*, 1733–1750.

Canziani, A., Culurciello, E., & Paszke, A. (2017). *An analysis of deep neural network models for practical applications*. arXiv:1605.07678v4.

Cohen, M. A., Alvarez, G. A., Nakayama, K., & Konkle, T. (2016). Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of Neurophysiology*, *117*, 388–402.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Proceedings of the European conference on computer vision* (pp. 1–22).

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.

Desimone, R., Schein, S., Moran, J., & Ungerleider, L. (1985). Contour, color and shape analysis beyond the striate cortex. *Vision Research*, *25*, 441–452.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*, 1827–1837.

DiCarlo, J., & Cox, D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*, 333–341.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14), 18–18.

Engel, S., Glover, G., & Wandell, B. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex*, *7*, 181–192.

Fize, D., Vandeffel, W., Nelissen, K., Denys, K., d'Hotel, C. C., Faugeras, O., & Orban, G. (2003). The retinotopic organization of primate dorsal v4 and surrounding areas: A functional magnetic resonance imaging study in awake monkeys. *The Journal of Neuroscience*, *23*, 7395–7406.

Freeman, J., & Simoncelli, E. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*, 1195–1201.

Gattas, R., Sousa, A., Mishkin, M., & Ungerleider, L. (1997). Cortical projections of area v2 in the macaque. *Cerebral Cortex*, *7*, 110–129.

Grill-Spector, K., Weiner, K. S., Gomez, J., Stigliani, A., & Natu, V. S. (2018). The functional neuroanatomy of face perception: From brain measurements to deep neural networks. *Interface Focus*, *8*, 20180013.

Harvey, B., & Dumoulin, S. (2011). The relationship between cortical magnification factor and population receptive field size in human visual cortex: Constancies in cortical architecture. *Journal of Neuroscience*, *31*, 13604–13612.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: surpassing human-level performance on imagenet classification In *Proceedings of the international conference on computer vision (CVPR)* (pp. 1026–1034).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR 2016)* (pp. 770–778).

Hong, H., Yamins, D., Majaj, N., & DiCarlo, J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*, 613–622.

Hout, M. C., Robbins, A., Godwin, H. J., Fitzsimmons, G., & Scarince, C. (2017). Categorical templates are more useful when features are consistent: Evidence from eye-movements during search for societally important vehicles. *Attention, Perception, & Psychophysics*, *79*, 1578–1592.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the international conference on computer vision (ICCV)*.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*.

Kastner, S., Weerd, P., Desimone, R., & Ungerleider, L. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science*, *282*, 108–111.

Kastner, S., Weerd, P., Pinsk, M., Elizondo, M., Desimone, R., & Ungerleider, L. (2001). Modulation of sensory suppression: Implications for receptive field sizes in the human visual cortex. *Journal of Neurophysiology*, *86*, 1398–1411.

Khaligh-Razavi, S., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*, e1003915.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. *Oxford Research Encyclopaedia of Neuroscience*. doi:10.1093/acrefore/9780190264086.013.46

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, *71*, 856–867.

Kravitz, D., Kadharbatcha, S., Baker, C., Ungerleider, L., & Mishkin, M. (2013). The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, *17*, 26–49.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modelling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105). Red Hook, NY: Curran Associates Inc.

Larsson, J., & Heeger, D. (2006). Two retinotopic visual areas in human lateral occipital cortex. *Journal of Neuroscience*, *26*, 13128–13142.

Li, M., & Tsien, J. Z. (2017). Neural code—neural self-information theory on how cell-assembly code rises from spike time and neuronal variability. *Frontiers in Cellular Neuroscience*, *11*, 236.

Li, M., Xie, K., Kuang, H., Liu, J., Wang, D., & Fox, G. (2017). *Spike-timing patterns conform to a gamma distribution with regional and cell type-specific characteristics*. BioRxiv:145813.

Li, G., & Yu, Y. (2015). Visual saliency based on multiscale deep features. In *IEEE conference on computer vision and pattern recognition (CVPR 2015)*.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*, 91–110.

Maxfield, J. T., Stalder, W. D., & Zelinsky, G. J. (2014). Effects of target typicality on categorical search. *Journal of Vision*, *14* (12), 1–11.

Maxfield, J. T., & Zelinsky, G. J. (2012). Searching through the hierarchy: How level of target categorization affects visual search. *Visual Cognition*, *20*(10), 1153–1163.

McKeefry, D., & Zeki, S. (1997). The position and topography of the human colour centre as revealed by functional magnetic resonance imaging. *Brain*, *120*, 2229–2242.

Mishkin, M., Ungerleider, L., & Macko, K. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, *6*, 414–417.

Nakamura, H., Gattass, R., Desimone, R., & Ungerleider, L. (1993). The modular organization of projections from areas v1 and v2 to areas v4 and teo in macaques. *The Journal of Neuroscience*, *13*, 3681–3691.

Nako, R., Wu, R., & Eimer, M. (2014). Rapid guidance of visual search by object categories. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 50–60.

Nassi, J. J., & Callaway, E. (2009). Parallel processing strategies of the primate visual system. *Nature Reviews Neuroscience*, *10*, 360–372.

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*, 614–621.

Nelson, W. W., & Loftus, G. R. (1980). The functional visual field during picture viewing. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(4), 391–399.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*, 145–175.

Orban, G., Zhu, Q., & Vanduffel, W. (2014). The transition in the ventral stream from feature to real-world entity representations. *Frontiers in Psychology*, *5*, 695.

Pasupathy, A., & Connor, C. (2002). Population coding of shape in area V4. *Nature Neuroscience*, *5*(12), 1332–1338.

Peelen, M. V., & Kastner, S. (2011). A neural basis for real-world visual search in human occipitotemporal cortex. *Proceedings of the National Academy of Sciences*, *108*(29), 12125–12130.

Rajimehr, R., Young, J., & Tootell, R. (2009). An anterior temporal face patch in human cortex, predicted by macaque maps. *Proceedings of the National Academy of Sciences*, *106*, 1995–2000.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *Computer vision and pattern recognition workshops*.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Rousselet, G., Thorpe, S., & Fabre-Thorpe, M. (2004). How parallel is visual processing in the ventral pathway? *Trends in Cognitive Sciences*, *8*, 363–370.

Russakovsky, O., Deng, J., Su, H., Jrause, J., Satheesh, S., Ma, S., … Li, F. F. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

Sanchez, J., Perronnin, F., Mensink, T., & Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, *105*, 222–245.

Schmidt, J., & Zelinsky, G. J. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quarterly Journal of Experimental Psychology*, *62*(10), 1904–1914.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress in Brain Research*, *165*, 33–56.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*, 6424–6429.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations (ICLR 2015)*.

Smith, A., Williams, A., & Greenlee, M. (2001). Estimating receptive field size from fMRI data in human striate and extrastriate visual cortex. *Cerebral Cortex*, *11*, 1182–1190.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., & Reed, S. (2015). Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition (CVPR 2015)*.

Tanaka, K. (1997). Mechanisms of visual object recognition: Monkey and human studies. *Current Opinion in Neurobiology*, *7*, 523–529.

Tarr, M. (1999). News on views: Pandemonium revisited. *Nature Neuroscience*, *2*, 932–935.

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bülthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience*, *14*, 869–876.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, *78*, 507–545.

Ungerleider, L., Galkin, T., Desimone, R., & Gattass, R. (2007). Cortical connections of area v4 in the macaque. *Cerebral Cortex*, *18*, 477–499.

Van Essen, D. C., Lewis, J., Drury, H., Hadjikhani, N., Tootell, R., Bakircioglu, M., & Miller, M. (2001). Mapping visual cortex in monkeys and humans using surface-based atlases. *Vision Research*, *41*, 1359–1378.

Vicente, T., Hoai, M., & Samaras, D. (2015). Leave-one-out kernel optimization for shadow detection In *Proceedings of the international conference on computer vision (ICCV)* (pp. 3388–3396).

Wade, A., Brewer, A., Rieger, J., & Wandell, B. (2002). Functional measurements of human ventral occipital cortex: Retinotopy

and colour. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *357*, 963–973.

Wang, W., & Shen, J. (2017). *Deep visual attention prediction*. arXiv:1705.02544.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, *1*, 202–238.

Yamins, D., Hong, H., Cadieu, C., Solomon, E., Seibert, D., & DiCarlo, J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*, 8619–8624.

Yang, H., & Zelinsky, G. J. (2009). Visual search is guided to categorically-defined targets. *Vision Research*, *49*, 2095–2103.

Yu, C.-P., Hua, W.-Y., Samaras, D., & Zelinsky, G. J. (2013). Modeling clutter perception using parametric proto-object partitioning. *Proceedings of the 26th Conference on Advances in Neural Information Processing Systems (NIPS 2013)*.

Yu, C.-P., Le, H., Zelinsky, G. Z., & Samaras, D. (2015). Efficient video segmentation using parametric graph partitioning. *International conference on computer vision (ICCV)*.

Yu, C.-P., Maxfield, J. T., & Zelinsky, G. J. (2016). Searching for category-consistent features: A computational approach to understanding visual category representation. *Psychological Science*, *27*(6), 870–884.

Zagoruyko, S., & Komodakis, N. (2016). *Wide residual networks*. arXiv preprint arXiv:1605.07146.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of the European conference on computer vision (ECCV 2014)*.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*(4), 787–835.

Zelinsky, G. J., Adeli, H., Peng, Y., & Samaras, D. (2013). Modelling eye movements in a categorical search task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368* (1628), 1–12.

Zelinsky, G. J., Peng, Y., Berg, A. C., & Samaras, D. (2013). Modeling guidance and recognition in categorical search: Bridging human and computer object detection. *Journal of Vision*, *13*(3), 30, 1–20.

Zelinsky, G. J., Peng, Y., & Samaras, D. (2013). Eye can read your mind: Using eye fixations to classify search targets. *Journal of Vision*, *13*(14), 10, 1–13.

Zhang, M., Feng, J., Ma, K. T., Lim, J. H., Zhao, Q., & Kreiman, G. (2018). Finding any Waldo: Zero-shot invariant and efficient visual search. *Nature Communications*, *9*, 3730.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014). *Object detectors emerge in deep scene CNNs*. arXiv:1412.6856.