Noname manuscript No.
(will be inserted by the editor)

Neural network models of object recognition can also account for visual search behavior

David A. Nicholson · Astrid A. Prinz

Received: date / Accepted: date

Abstract What limits our ability to find an object we are looking for? There are two competing models: one explains attentional limitations during visual search in terms of a serial processing computation, the other attributes limitations to noisy parallel processing. Both models predict human visual search behavior when applied to the simplified stimuli often used in experiments, but it remains unclear how to extend them to account for search of complex natural scenes. Models exist of natural scene search, but they do not predict whether a given scene will limit search accuracy. Here we propose an alternate mechanism to explain limitations across stimuli types: visual search is limited by an "untangling" computation, proposed to underlie object recognition. To test this idea, we ask whether models of object recognition account for visual search behavior. The current best-in-class models are artificial neural networks (ANNs) that accurately predict both behavior and neural activity in the primate visual system during object recognition tasks. Unlike dominant visual search models, ANNs can provide predictions for any image. First we test ANN-based object recognition models with simplified stimuli typically used in studies of visual search. We find these models exhibit a hallmark effect of such studies: a drop in target detection accuracy as the number of distractors increases. Further experiments show this effect results from learned representations: networks that are not pre-trained for object recognition can achieve near perfect accuracy.

Research funded by the Lifelong Learning Machines program, DARPA/Microsystems Technology Office, DARPA cooperative agreement HR0011-18-2-0019. David Nicholson was partially supported by the 2017 William K. and Katherine W. Estes Fund to F. Pestilli, R. Goldstone and L. Smith, Indiana University Bloomington.

David A. Nicholson
Emory University
Department of Biology,
O. Wayne Rollins Research Center,
1510 Clifton Road NE, Atlanta , GA 30322 E-mail: nicholdav@gmail.com

Astrid A. Prinz
Emory University
Department of Biology,
O. Wayne Rollins Research Center,
1510 Clifton Road NE, Atlanta , GA 30322

Next we test these models with complex natural images, using a version of the Pascal VOC dataset where each image has a visual search difficulty score, derived from human reaction times. We find models exhibit a drop in accuracy as search difficulty score increases. We conclude that ANN-based object recognition models account for aspects of visual search behavior across stimuli types, and discuss how to extend these results.

Keywords Visual search · Selective visual attention · Object recognition · neural networks · Deep learning

1 Introduction

What limits our ability to find an object as we search for it in the cluttered scenes constantly presented to our visual system? Essentially, there are two competing families of models that explain limitations on visual search, as schematized in Fig. 1. One family attributes limitations to a serial processing step required to bind different feature dimensions like color and orientation into a target object (Wolfe, 1994; Wolfe et al., 1989; Treisman and Gelade, 1980). The other family attributes limitations to noisy parallel processing (Eckstein, 1998; Eckstein et al., 2000; Palmer et al., 2000; Eckstein, 2011). These have also been described as models of visual selective attention (Eckstein, 1998; Geisler and Cormack, 2011; Wolfe and Horowitz, 2017; Peelen and Kastner, 2014). However, we defer usage of the term "attention" until the discussion, to place emphasis on which specific computations impose limits on search behavior (Hommel et al., 2019), and to avoid overloading the term, since "attention" has a different meaning when applied to the artificial neural network models that we test here (Lindsay, 2020; Hommel et al., 2019). The dominant models explaining limitations are built on results obtained by measuring search behavior with highly simplified stimuli like those shown in the upper left of Fig. 1, that have been employed in thousands of studies. As can be seen, the stimuli consist of a two-dimensional array of items, where a target is usually distinguished from distractors by one or two parametrically-defined features like hue, luminance, or orientation. Accordingly, the dominant models accurately predict visual search behavior when given as inputs some number of items that are described by a few clearly-defined features. We emphasize that what both families of models have in common is that they assert that the brain operates on well-defined features like color and orientation that it uses while processing sets of discrete items.

These assertions shared by both families of models make it unclear how to extend them to relatively complex natural scenes, more similar to what we see during real world search, like those shown in the upper right of Fig. 1. In real world scenes, an object we are searching for will not differ from nontarget items by one or two parametrically-defined features like color and orientation (Peelen and Kastner, 2014; Wolfe et al., 2011). In addition, natural scenes are not easily divided into discrete items (Hulleman and Olivers, 2017). Even if it were possible to easily extend existing models to natural images, their predictions might not match real world visual search behavior. For example, models with a serial processing step imply that search of cluttered scenes should produce lengthy reaction times, but often these searches can be highly efficient (Peelen and Kastner, 2014; Wolfe

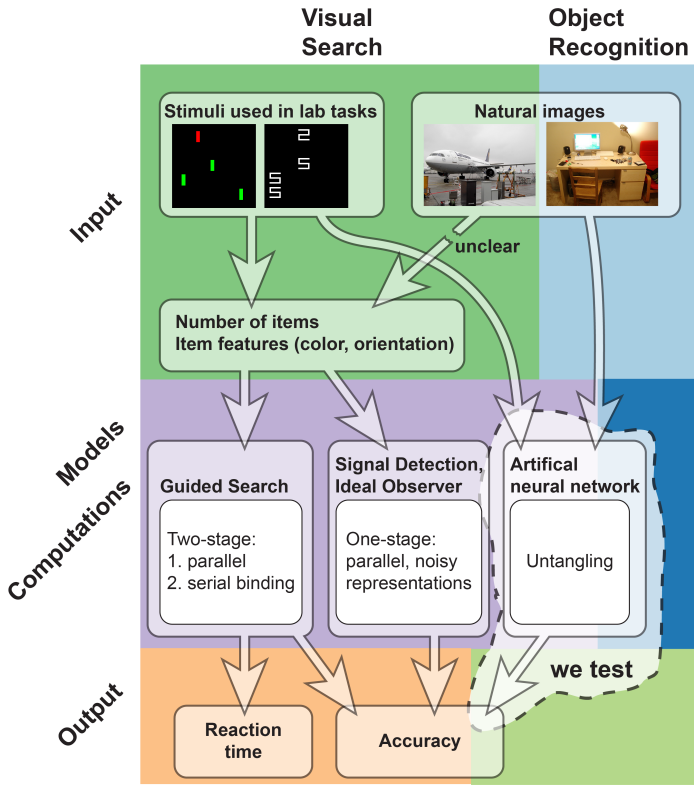

Fig. 1 Schematic representation of hypothesis we test, and its relationship to other models. See text for description

et al., 2011). Finally, dominant models of visual search are not consistent with our current understanding of the visual system (Nakayama and Martini, 2011). One way in which this is true is the emphasis those models place on what features guide search (Wolfe and Horowitz, 2017). Emerging evidence from neuroscience suggests that, beyond the sensory periphery, neural activity in the visual system is not dominated by neurons representing or encoding separate feature dimensions like color or orientation (Nakayama and Martini, 2011), especially during natural behavior (de Vries et al., 2020).

Models exist that explain visual search of natural scenes, but they do not directly account for limitations of search performance in terms of a specific computation. A conceptual model has been described based on proposed functions of the dorsal and ventral pathways of the visual system (Peelen and Kastner, 2014), and imaging studies provide evidence in support of this model, but to our knowledge no computational studies of the model have been carried out. Conversely, experiments have shown that that visual search performance can be predicted from features of natural scenes (Katti et al., 2017; Ionescu et al., 2016), but this predictive power of features was not attributed to a specific computation carried out by the visual system.

Here we ask whether a specific computation can account for limitations on visual search behavior across stimuli. Specifically, we hypothesize that the core computation proposed to underlie object recognition is also required for visual search, and that it places limits on visual search performance. The key computation proposed to underlie object recognition is "untangling": transforming low-level features of the visual scene into a representational space where the object can be classified (DiCarlo and Cox, 2007; DiCarlo et al., 2012). According to this theory, the visual system learns to project different views of an object onto a low-dimensional manifold in "object space", such that a hyperplane exists separating that manifold from other object manifolds. Several previous studies provided evidence suggesting that a similar computation might be involved in visual search (Bauer et al., 1996; Duncan and Humphreys, 1989; D'Zmura, 1991), as noted by (Nakayama and Martini, 2011). Those previous studies used the simplified stimuli described above, and the authors attributed behavioral effects to the intrinsic discriminability of target and distractors. In contrast, our central hypothesis is that untangling itself can account for limitations on target detection accuracy that are observed when human subjects search those simplified stimuli, and additionally can account for limitations on accuracy seen when human subjects search natural scenes.

To test our hypothesis, we use artificial neural networks (ANNs) that are currently best-in-class models of object recognition (Yamins and DiCarlo, 2016; Schrimpf et al., 2018). These ANN models accurately predict both neural activity in the primate visual system and behavior during object recognition tasks. Specifically, ANN accuracy is correlated with accuracy of human subjects and monkeys performing object recognition tasks. The models are optimized for task performance, in this case image classification, using methods that have come to be called deep learning (Yamins and DiCarlo, 2016; Yamins et al., 2014; Richards et al., 2019). An additional benefit of using ANNs is that they are image-computable models (Yamins et al., 2014; Geisler and Cormack, 2011), meaning they can provide predictions both for the highly simplified stimuli used in lab tasks and for natural scenes.

Previous studies have tested different ANN architectures as models of visual search (Poder, 2017; Eckstein et al., 2017; Grossberg et al., ????; Ma et al., 2011), but we emphasize that here our goal is to specifically test whether ANN-based object recognition models exhibit human-like behaviors when performing visual search tasks. If so, we will take this as evidence that the untangling computation proposed to underlie object recognition also represents an alternate mechanism that could account for visual search behavior. To observe the behavior of ANN-based object recognition models performing visual search task, we use methods from deep learning to adapt pre-trained models to new domains. As we show below, this approach produced results supporting the idea that an untangling computation can account for limitations of target detection accuracy across the different types of stimuli used in visual search tasks.

2 General Methods

2.1 Transfer learning

For all experiments, we employ transfer learning methods used with deep neural network models (Yosinski et al., 2014; Kornblith et al., 2019). Essentially, we hold fixed all parameters optimized for object recognition, except for those in the final output layer of the neural network. We replace the final layer used for image classification with a new layer that has an appropriate number of units for the visual search task and then adapt the model to this task by optimizing for performance with a training set. We then measure visual search task performance with a held-out test set the network has not seen during training time.

2.2 Code availability

To aid with reproducibility of our experiments, and to make them more accessible to other researchers, we developed a separate software library, `visual-search-nets`, available at https://github.com/NickleDave/visual-search-nets. All code related to carrying out experiments, analyzing and visualization of results, is available at https://github.com/NickleDave/untangling-visual-search.

3 Results

In order to use ANN-based models to test the idea that optimizing the visual system for object recognition places limits on visual search behavior, we employ transfer learning methods used with deep neural network models (Yosinski et al., 2014; Kornblith et al., 2019). Essentially, we hold fixed all neural network parameters optimized for object recognition, except for those in the final output layer. We replace the final layer used for image classification with a new layer that has an appropriate number of units for the visual search task and then adapt the model to this task by optimizing for performance with a training set. For standard lab tasks using simplified stimuli, the final layer has two output units corresponding to "target present" and "target absent"; for tasks using natural scenes, the number of outputs units corresponds to the number of classes in the dataset (20 in the case of the Pascal VOC dataset we use). We then measure visual search task performance with a held-out test set the network has not seen during training time. For all experiments we test the performance of four ANN models. Two of the models, AlexNet (Krizhevsky et al., 2012) and VGG16 (Simonyan and Zisserman, 2015), represented key advances in image classification by the computer vision community, and were later used in some of the first papers that proposed ANNs as models of object recognition (Cadieu et al., 2014; Jozwik et al., 2017). The other two, CORnet S and CORnet Z, are two ANNs specifically developed in pursuit of good performance under a metric proposed to account for a model's ability to predict both brain activity and behavior during object recognition tasks (Schrimpf et al., 2018). Other models might be added but we chose these four architectures as a representative sample of ANN-based object recognition models

to increase the likelihood that our results are general, and are not an artifact of any specific architecture.

3.1 ANN models of object recognition exhibit set size effects

3.1.1 Set size effects are a hallmark of search tasks that use simplified stimuli

We first tested whether ANNs exhibited behavior similar to that seen when humans search highly-simplified stimuli like those shown in Fig. 2. The prevailing models of visual search are built on results obtained with these stimuli, which have been employed in thousands of studies. A complete review of the dominant models is beyond the scope of this paper, but there are essentially two competing families (Wolfe, 1994; Eckstein, 2011). The evidence for both families of models consists of so-called "set size effects", observed when subjects perform visual search tasks that use stimuli like those in Fig. 2. Since we test whether ANN-based object recognition models also exhibit set size effects, we define this term by briefly describing how these laboratory search tasks are typically performed (Wolfe, 1998). On each trial, the subject reports whether a target is present (Fig. 2**a** bottom row) or absent (Fig. 2**a** top row) among distractors. Studies using these stimuli use the term "set size" to refer to the total number of items (distractors plus target when present). Hence by extension the term "set size effect" refers to any change in some behavioral measure of target detection, such as reaction time or accuracy, that depends on increasing the number of distractors, as depicted schematically in Fig. 2**b**. These set size effects are taken as evidence for different types of computations thought to be involved in visual search. Therefore, we tested whether ANN models of object recognition exhibit set size effects, supporting the idea that the untangling computation can account for visual search. Specifically, we asked whether ANNs showed a decrease in accuracy of detecting a target as the number of distractors increased.

3.1.2 Methods

All ANN models were trained on a dataset consisting of ten types of stimuli (columns in Fig. 3), with 1200 samples for each type. Stimuli were generated with jitter in the placement of the items, to guarantee that there were no repeated images that would encourage the ANNs to simply memorize the correct answer during training. 1200 samples was the maximum number we could generate per stimulus without repeats, given the parameters we used to create them.

3.1.3 ANNs pre-trained for object recognition show set size effects, while ANNs trained from randomly initialized weights achieve perfect accuracy

As shown in Fig. 3, all ANN models we tested exhibited set size effects. We tested four different ANN architectures that have been used as models of object recognition (rows in Fig. 3), and all showed a drop in target detection accuracy as the number of distractors increased (solid blue line). This set size effect was consistent across the eight training replicates we ran for each model (dashed lines). All replicates were trained on a dataset consisting of ten types of stimuli (columns

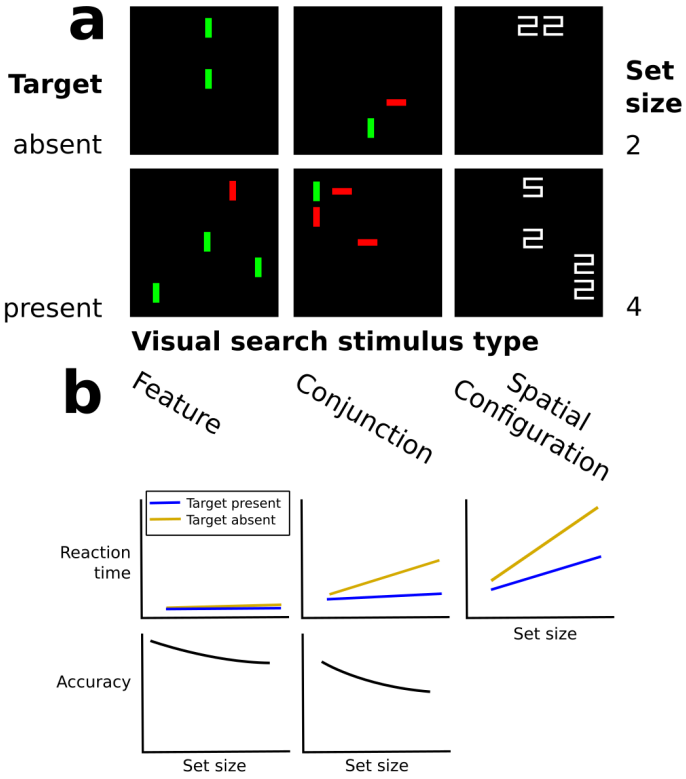

Fig. 2 Set size effects are a hallmark finding from laboratory visual search tasks. Panel **a** shows example simplified displays commonly used in visual search tasks. In top row of **a** a target is absent and in the bottom row it is absent. Displays in each row also have different set sizes (total number of items including target and distractors): on the top row of **a** the set size is two and the bottom row it is four. Panel **b** schematically depicts *set size effects*, redrawn from (Wolfe et al., 2010) and (Eckstein, 1998)). Effect size varies based on the features that distinguish targets from distractors (shown in columns). In the left column of **a**, the target can be distinguished from distractors by a single *feature*, color; in the middle column, by a *conjunction* of features, color and orientation; in the right column, by a *spatial configuration* of multiple features.

in Fig. 3) with 1200 samples for each type (see methods for details). The effect size was different for different stimulus types; we measured effect size by taking the difference between accuracy for set size 8 and accuracy for set size 1. The columns in Fig. 3 are sorted by the size of this effect, in increasing order, averaged across models. Note that when sorted this way it can be seen that stimulus types thought to give rise to "efficient" search in humans were easily discriminable by the networks, like the blue x target vs. red distractors in the last column in figure 2. Similarly, stimulus types thought to give rise to inefficient search, such as the "digital 2 target vs. digital 5 distractors stimulus in the first column, were difficult for the networks to discriminate.

This initial finding demonstrated that ANN models exhibit set size effects, but it left unanswered the question of whether these effects result from the models

being optimized for object recognition, or whether alternatively they result from the structure of ANNs in general. To answer this question, we carried out another set of experiments where we trained models from randomly-initialized weights, instead of using weights pre-trained on ImageNet. In almost all cases, these models achieve nearly perfect accuracy across all set sizes (solid red lines in fig. 3). For all models we tested, most of the eight replicates (dashed red lines) achieved these high levels of accuracies, providing an existence proof that these ANN models can find near-optimal solutions to this task. The only exceptions were one replicate of AlexNet that failed to converge, and one replicate of VGG16 that did not achieve high accuracy on two of the ten stimulus types.

A final concern that might be raised about our results is that the simplified stimuli we used somehow affect the activations within the hidden layers of the neural networks, impeding their ability to learn the task. For example, the black backgrounds might produce lower activations on average than the activations produced by full-color images from ImageNet used to pre-train the model. To address this concern, we carried out a control experiment where we produced the same set of simplified stimuli types, only with a white background instead of black, and we repeated the training with the AlexNet model. We again saw that AlexNet models pre-trained on ImageNet exhibited set size effects, whereas AlexNet models trained from randomly-initialized weights were able to achieve very high accuracy on the same task. Please see the repository on-line for these results. Thus this control experiment did not produce evidence that our result is an artifact of the stimulus.

Based on the results from this set of experiments, we conclude it is possible to optimize ANNs to perform this task with perfect accuracy, when parameters are not pre-trained for image classification. Therefore the set size effects we observed resulted from ANNs being optimized for object recognition before we used transfer learning to measure how they performed visual search tasks.

3.1.4 set size effects exhibited by ANN models do not arise from overfitting, failure of optimization to converge, or lack of training data

We found that ANNs pre-trained for image classification showed a drop in accuracy of target detection as the number of distractors increased, as shown in Fig. 3. An alternative explanation for our findings would be that we simply failed to find the best method to fit models. The method we use here is based on preliminary experiments (Nicholson and Prinz, 2019) where we eliminated the possibility that set size effects arose from: hyperparameters such as learning rate; imbalance in the data set; and the size of the training set. In addition, we took several steps to minimize the possibility that the results presented here were an artifact of our training method. Those steps included logging metrics at each step of training, then visually assessing plots of the logged training histories for evidence of overfitting to the training set, or failure of the optimization to converge. In Fig. 4 we show representative training histories from the Alexnet model. (Plots of training histories for all models can be found in the openly shared repository accompanying this paper; see methods for link.) In almost all cases, we saw that the loss function converged to an asymptotic value on the training set, as well as on a validation set that ANN models did not see during training. We also saw that the models achieved high accuracy on this validation set, indicating that what they learned during training generalized to unseen data. The only exception was one replicate

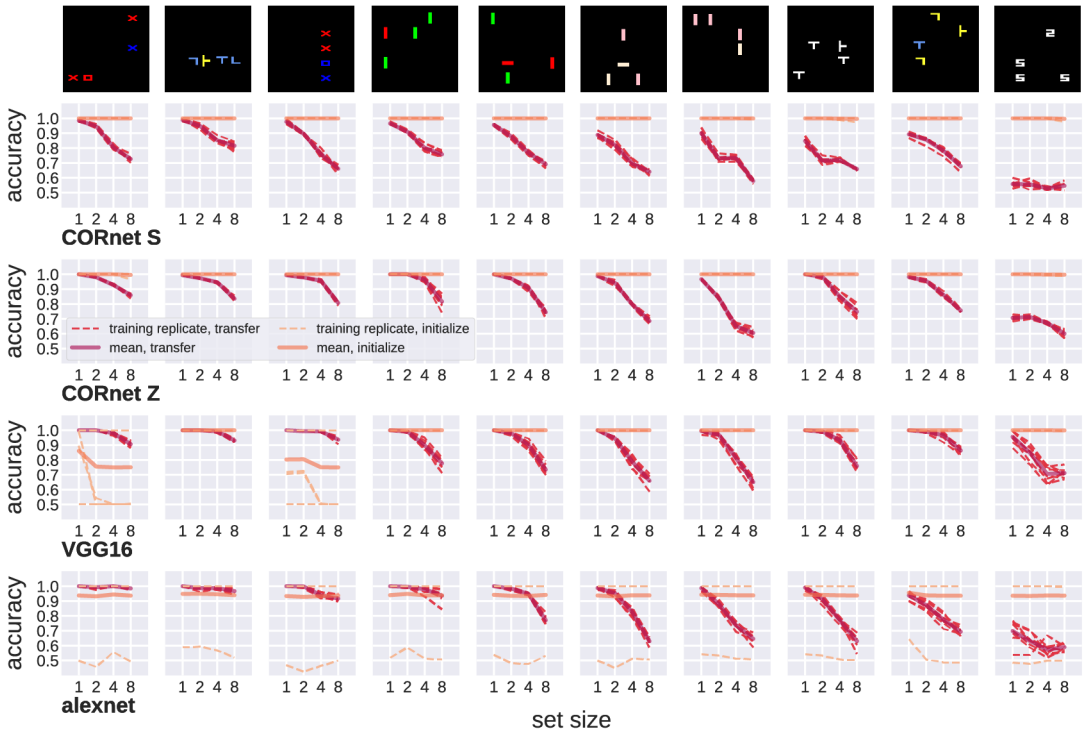

Fig. 3 ANNs pre-trained for image classification show set size effects, while ANNs trained from randomly initialized weights can achieve near perfect accuracy. Rows show four different models trained on datasets consisting of 10 different search stimulus types (examples shown at top of columns). Each axes shows accuracy as a function of set size, for eight training replicates (dashed lines). Solid lines indicate mean across all trials and replicates. Brown dashed lines and grey solid lines indicate accuracy for object recognition models, that had weights pre-trained for image classification and were adapted to this visual search task using transfer learning. Peach dashed lines and salmon solid lines indicate accuracy for models trained from randomly-initialized weights, not pre-trained for image classification. Columns are ordered by effect size, the difference in accuracy between set size 1 and 8 for the object recognition models. Rows are organized in increasing order of per-model effect size, averaged across all stimulus types. All example images for different stimulus types are shown with the target present condition. Stimulus types from left to right are: blue x target v. red x and red o distractors; yellow T target v. blue T and blue L distractors; blue x target v. red x and blue o distractors; red vertical line target v. green vertical line distractors; red vertical line target v. red horizontal and green vertical line distractors; papaya whip vertical line target v. papaya whip horizontal and pink vertical distractors; papaya whip vertical line target v. pink vertical distractors; rotated T target v. T distractors; yellow rotated T target v. blue T and yellow L distractors; white digital 2 target v. white digital 5 distractors

of one model that failed to converge. Hence, we do not find any evidence that set size effects can be attributed to an artifact of training.

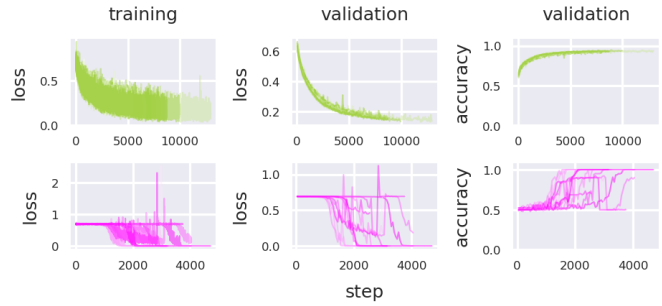

Fig. 4 Set size effects exhibited by ANN models do not arise from overfitting or failure of optimization to converge. Shown are representative training histories from the Alexnet model. (Plots of training histories for all models can be found in the openly shared repository accompanying this paper; see methods for link.) Top row, training histories from experiments using transfer learning to hold parameters fixed that were pre-trained for object recognition; bottom row, histories for the same model trained "from scratch" with parameters randomly initialized. For each ANN architecture and training method (transfer or from randomly initialization), we trained eight replicates, each indicated by a different shade of color. For all models and methods, we saw that the loss function converged to an asymptotic value on the training set (column 1), as well as on a validation set that ANN models did not see during training (column 2). In addition, the models achieved high accuracy on this validation set (columns 3), indicating that what they learned during training generalized to unseen data.

3.2 Accuracy of ANNs correlates with visual search difficulty scores based on reaction times of humans searching natural scenes

As described in the introduction, a key limitation of dominant models of visual search is that it is unclear how to extend them to account for search of relatively complex natural scenes. We propose that the untangling computation provides an alternative account of search behavior across stimulus types, and leverage the ability off ANNs to provide predictions for any type of image to test our proposition. A key finding from experiments using natural scenes to study visual search is that it can be highly efficient, as suggested by reaction times. In spite of this efficiency, different natural scenes elicit different reaction times, as evidenced by previous attempts to estimate the extent to which targets and contextual information influence real world visual search (Katti et al., 2017; Ionescu et al., 2016). Although the ANN models we test here do not directly produce reaction times as an output, we can ask whether their accuracy correlates with reaction times.

3.2.1 The Visual Search Difficulty dataset

To test for a relationship between ANN accuracy and human reaction times during visual search, we use the Visual Search Difficulty dataset, first presented in (Ionescu et al., 2016). The dataset consists of search difficulty scores assigned to all images in the Pascal VOC dataset. These scores are a weighted combination of reaction times of human subjects asked to report whether a target is present in the image (please see (Ionescu et al., 2016) for details). In Fig. 5 we present representative images and scores from the dataset (top row), as well as summary statistics (bottom row). While we cannot directly test whether ANN outputs pre-

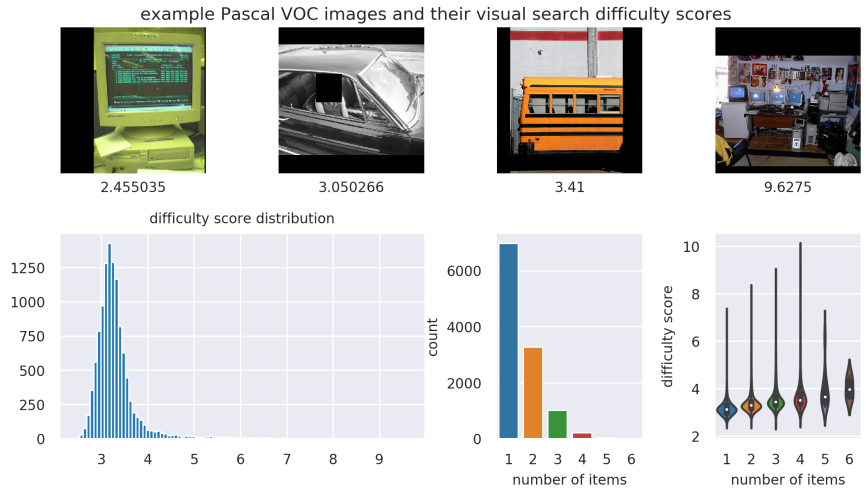

Fig. 5 The Visual Search Difficulty dataset assigns search difficulty scores to images from the Pascal VOC dataset, based on reaction times of human subjects searching each image for targets In the top row, example images from the Pascal VOC dataset are shown, and the visual search difficulty score assigned to that image is shown underneath it. Images are the first through fourth quartile of the distribution. Face in second image obscured for privacy. The distribution of scores is shown in the bottom panel of the first row. Note the long tail of the distribution: the median around 3 but that scores extend all the way to 9. In terms of the number of items, most images have only one item (annotated), but there are a good number with two or three. In the last panel we see that a slight trend where images with more items have a higher difficulty score.

dict human behavior using the dataset, we can ask if they are related, as a first step towards testing whether untangling can account for real world visual search behavior.

3.2.2 Target detection accuracy of ANN models of object recognition correlates with visual search difficulty scores

We do find a clear relationship between accuracy of ANN models and search difficulty scores based on human reaction times, as shown in Fig. 6. Because we could not directly link model outputs to reaction times, we chose to bin the difficulty scores, then measure accuracy of each trained model on the images within each bin. To do so, we binned the difficulty scores such that there were equal numbers of items in each bin, so that estimates of accuracy were based on similar sample sizes across bins. This analysis revealed a strong correlation between accuracy and difficulty score: as shown in Fig. 6, for all models, accuracy dropped as the search difficulty score increased. Because there are multiple objects present in the images in this dataset, we tested for this relationship in two ways. First, we adapted models using transfer learning methods as before, where each model classifies an image with a single label. We trained models this way using either the largest object in an image, as defined by its bounding box, or a randomly chosen image. The object chosen had no noticeable effect on accuracy when measured on a test set. We also carried out a separate set of experiments where we trained models

for multi-label classification, meaning that that model could classify an image as having any number of objects from the twenty classes present in the dataset. In both cases, the relationship between accuracy and visual search difficulty score held. Based on these results, we conclude that optimizing ANN models for object recognition limits their ability to perform a search task using complex natural images.

In experiment one we showed that models trained from randomly-initialized weights were able to achieve near-perfect accuracy when classifying simplified search stimuli as "target present" or "target absent". One might reasonably ask why not carry out the same experiment here. For this experiment with natural images, we would expect that models trained from randomly-initialized weights would not achieve higher accuracy than models with weights pre-trained for image classification on ImageNet, because transfer learning is known to provide better accuracy than training "from scratch". In particular this has been shown when pre-training models on ImageNet and then adapting the models to a related task where we have a much smaller training set (Yosinski et al., 2014; Kornblith et al., 2019), as is the case here with the PascalVOC dataset used to derive visual search difficulty scores. For the sake of completeness, we did train models from randomly-initialized weights using single- and multi-label classification, and did find that those models achieved much lower accuracy on the test set. Please see on-line repository for results. It could be argued that in this case we are simply training models to perform object recognition, not a visual search task. We return to this point in the discussion. Because models trained from randomly-initialized weights did not achieve high accuracy on the test set, we did not analyze their performance in terms of visual search difficulty scores.

4 Discussion

We set out to test whether the untangling mechanism proposed to underlie object recognition can also account for limitations on visual search performance across visual search tasks. Our goal was to address unreconciled findings in the literature, where visual search behavior obtained with simplified stimuli has been used to support models that are not easily extended to account for search of relatively complex natural scenes. To test whether untangling can account for search behavior across simplified stimuli and relatively complex natural scenes, we used ANN-based models object recognition models that capture the key computation of untangling. A key advantage of ANN models, compared to prevailing models of visual search, is that they are image-computable, i.e. they can produce predictions for any image. This characteristic of ANN models allowed us to to measure their behavior with two types of stimuli commonly used in lab tasks, the simplified displays shown in Fig. 2 and natural scenes like those shown in Fig. 5. Our results demonstrate that ANN-based object recognition models exhibit behaviors characteristic of humans performing visual search tasks. Unlike prevailing models of visual search, ANN models can produce predictions for any image, i.e. they are image-computable. When tested with simplified displays, the ANN models exhibited set size effects that are hallmarks of this task, as shown in Fig. 3. Similarly, when detecting targets in natural scenes, ANN models also a drop in accuracy inversely correlated with visual search difficulty scores derived from reaction times of

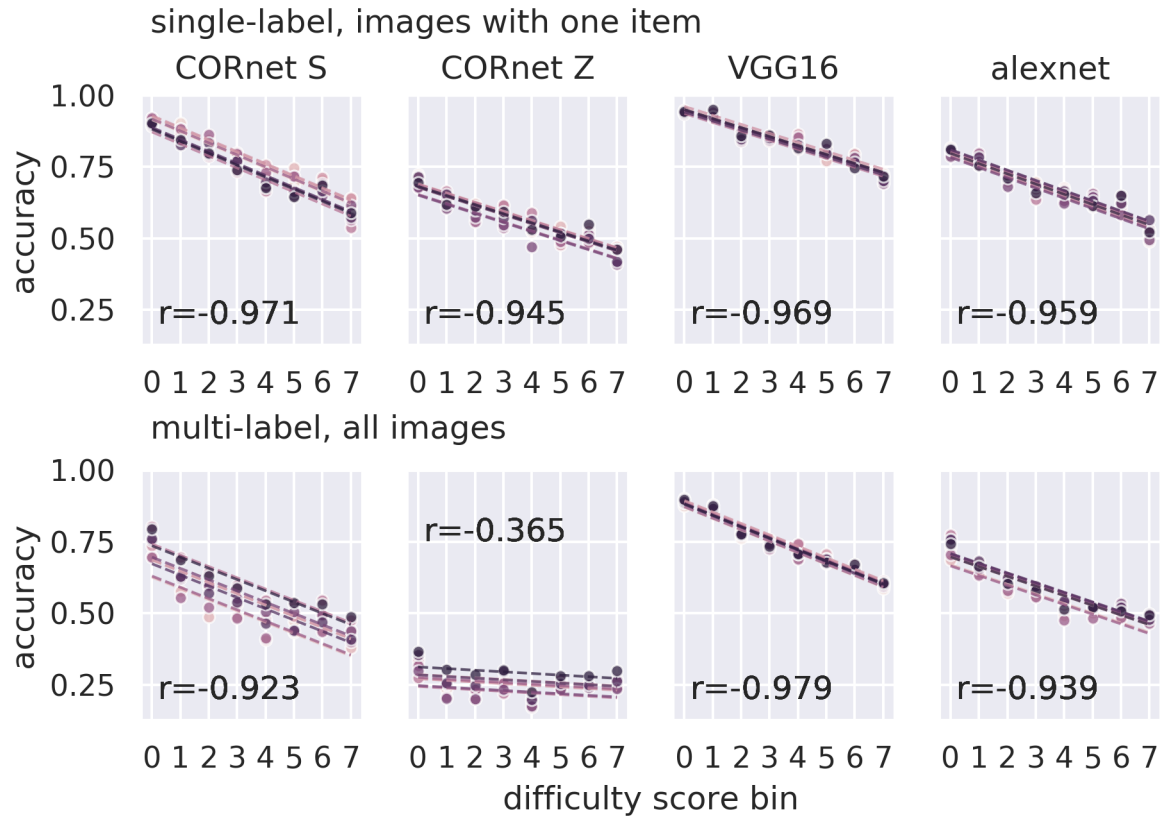

Fig. 6 Target detection accuracy of ANN models of object recognition correlates with visual search difficulty scores.
We observed a clear correlation between binned visual search difficulty scores (x axes) and accuracy of ANN-based object recognition models when measured on images within a given bin (y axes). This was true for all models (columns) when measuring accuracy using images where only a single item was labeled (top row). The effect was consistent across all eight training replicates (dashed lines with different shades). We saw a similar effect when adapting the ANNs pre-trained for object recognition to predict multiple labels for images (bottom row). In this case all models except the CORnet Z model showed a strong negative correlation between search difficutly score and accuracy.

human subjects searching the images, as shown in Fig. 6. To further test whether the set size effects we saw with simplified stimuli were specific to object recognition models, we carried out separate experiments where we trained the same ANN architectures from randomly-initialized weights, instead of using weights pre-trained for image classification, as is typically done for object recognition models (Yamins et al., 2014; Schrimpf et al., 2018). This experiment demonstrated that models trained from randomly-initialized weights are capable of performing the task with near-perfect accuracy, as we show in Fig. 3, and so we conclude that the effect is specific to object recognition models. We conclude that taken in whole our results support our assertion that untangling provides an alternative mechanism that can account for limitations on visual search performance as typically measured with

lab tasks. Hence our findings point to a mechanism that could potentially resolve findings across the literature. Of course this will need to be tested empirically by more direct comparisons of ANN model behavior with that of human subjects.

This need to further test empirically points to one caveat of our findings. We feel the studies here are complete, because they can can account for limitations on visual search across stimuli in terms of one commonly used behavioral measure: accuracy. However, the models we test here cannot account for other behavioral measures, the most crucial of which is reaction time. This is because the ANN models we employ do not explicitly represent time. Previous studies have addressed this limitation in a couple of ways. One is to limit the time that subjects can view an image, to say 100 milliseconds, and compare accuracy in this time-limited setting to the accuracy of the ANN models. In part, this approach has been justified by the need to compare activation in the ANN models with neural activity in some fixed window. For extensive discussion of this operationalization of object recognition see (Cadieu et al., 2014; Majaj et al., 2015) and references therein. The time-limited 100 ms stimulus presentation is most relevant to our findings. Researchers studying visual search have also often carried out experiments where they limit the time that subjects view a stimulus, although in that case this experimental design is motivated by the need to control other factors such as eye movement (Eckstein, 1998; Wolfe, 1998; Eckstein, 2011). In other words, our approach for modeling visual search aligns with a standard approach for modeling object recognition, although for different reasons. At the least our approach makes sense given that researchers using the same restrictions on human subjects performing visual search tasks. As an alternative to studying accuracy of object recognition in a time-limited setting, some studies have used recurrent neural networks, which carry out a computation for a specified number of time steps t, in some cases to specifically model object recognition tasks without time constraints (Kar et al., 2019; Kietzmann et al., 2019; Spoerer et al., 2017; Nayebi et al., ????). In general these studies find that recurrence conveys an advantage in terms of predicting neural activity and behavior. However it is not clear how to relate this implicit time step to clock time in seconds, which would be needed for models to directly predict reaction time. A more direct solution would be to add a computation to our models that endow their behavior with reaction times, such as a winner-take-all mechanism for decision making. One previous computational study (Narbutas et al., 2017) did succeed in directly comparing reaction times of visual search models limited by serial binding (Moran et al., 2013) to models limited by noisy parallel processing, by adding a winner-take-all mechanism to the noisy parallel model. Although the same mechanism could be applied to ANN models of object recognition so that they produce a reaction time, the models would always produce the same reaction time given a particular image, since their output is deterministic (at least, at inference time, ignoring things like stochastic dropout often used during training). In contrast, human subjects produce a distribution of reaction times across trials (Wolfe et al., 2010).

Given these caveats, we suggest two lines of future work. The first approach we suggest that would address questions abouty reaction time would be to adopt a modeling framework that makes it possible to combine ANN models tested here with higher-level cognitive models. For example, the Neural Engineering Framework (Eliasmith and Anderson, 2003; Eliasmith and Stewart, ????) for cognitive modeling has at its core a "neural compiler" (Bekolay et al., 2014) that converts

cognitive abilities specified as functions into spiking neural networks, and the recently developed Nengo-DL library (Rasmussen, 2019) makes it possible to integrate deep neural networks like those we study here into Nengo models. Previous work with Nengo has considered families of winner-take-all mechanisms (Gosmann et al., ????) making it possible to easily endow the ANN models studied here with reaction times, and thus ask whether models explain behavior across different tasks. For example, future work could test which if any of those winner-take-all mechanisms produce distributions of reaction times that match human subjects.

Another line of future work we suggest would be to integrate the study of object recognition and visual search. A cynic might argue that our study simply shoehorns visual search tasks into the modeling framework used to study object recognition. We do agree that the tasks we use here to study visual search are in many respects similar to tasks used to study object recognition, and we feel that this emphasize the need to better understand how those two behaviors relate to each other in the real world. One possibility, suggested by our findings, is that the primate visual system is optimized for object recognition, and that optimizing for this behavior places limits on visual search, a related behavior. In this way, our experiments can be said to adopt the recently proposed deep learning framework for neuroscience (Richards et al., 2019) to ask how optimizing for one behavior impacts others. As other authors have noted (Kell and McDermott, 2019), findings within this framework can be related to ideal observer models, which have been applied to visual search (Geisler, 2003). Ideal observer models take a normative approach, proposing a closed-form optimal solution for a task and then asking how real-world behavior deviates from the behavior expected from the optimal solution. Obviously there is no known closed-form solution for object recognition, but as has been noted previously, the optimization perspective at least provides empirical evidence that our models perform the task very well, at least as measured with the test data set. Future work could further test this idea, or alternatively ask whether a different objective function could produce models that better account for both object recognition and visual search behavior Saxe et al. (????).

5 Conclusion

In spite of any questions about the differences between tasks used to investigate object recognition and visual search behavior, we feel it is likely that untangling will remain a key mechanism in models that account for both visual search behavior and object recognition in the real world. Most neuroscientists will agree that the visual system (and the brain) performs a series of nonlinear transformations on its inputs, regardless of what behavior we engage in, and regardless of the nuances about which parts of the visual system can be mapped to which type of behavior. Put this way, it is in some sense trivial to argue that untangling can account for aspect of both behaviors. However the prevailing models of visual search do not account for these nonlinear transformations, and so we hold that the results here set forth a strong alternative mechanism to account for limitations on visual search accuracy across stimuli types.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Bauer B, Jolicoeur P, Cowan WB (1996) Visual search for colour targets that are or are not linearly separable from distractors. Vision research 36(10):1439–1466

Bekolay T, Bergstra J, Hunsberger E, DeWolf T, Stewart TC, Rasmussen D, Choo X, Voelker AR, Eliasmith C (2014) Nengo: A Python tool for building large-scale functional brain models. Frontiers in Neuroinformatics 7, DOI 10.3389/fninf.2013.00048

Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ (2014) Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. PLoS Computational Biology 10(12):e1003963, DOI 10.1371/journal.pcbi.1003963

de Vries SEJ, Lecoq JA, Buice MA, Groblewski PA, Ocker GK, Oliver M, Feng D, Cain N, Ledochowitsch P, Millman D, Roll K, Garrett M, Keenan T, Kuan L, Mihalas S, Olsen S, Thompson C, Wakeman W, Waters J, Williams D, Barber C, Berbesque N, Blanchard B, Bowles N, Caldejon SD, Casal L, Cho A, Cross S, Dang C, Dolbeare T, Edwards M, Galbraith J, Gaudreault N, Gilbert TL, Griffin F, Hargrave P, Howard R, Huang L, Jewell S, Keller N, Knoblich U, Larkin JD, Larsen R, Lau C, Lee E, Lee F, Leon A, Li L, Long F, Luviano J, Mace K, Nguyen T, Perkins J, Robertson M, Seid S, Shea-Brown E, Shi J, Sjoquist N, Slaughterbeck C, Sullivan D, Valenza R, White C, Williford A, Witten DM, Zhuang J, Zeng H, Farrell C, Ng L, Bernard A, Phillips JW, Reid RC, Koch C (2020) A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. Nature Neuroscience 23(1):138–151, DOI 10.1038/s41593-019-0550-9

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends in cognitive sciences 11(8):333–341

DiCarlo JJ, Zoccolan D, Rust NC (2012) How Does the Brain Solve Visual Object Recognition? Neuron 73(3):415–434, DOI 10.1016/j.neuron.2012.01.010

Duncan J, Humphreys GW (1989) Visual Search and Stimulus Similarity 96(3):433–458

D'Zmura M (1991) Color in visual search. Vision research 31(6):951–966

Eckstein MP (1998) The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. Psychological Science 9(2):111–118

Eckstein MP (2011) Visual search: A retrospective. Journal of vision 11(5):14–14

Eckstein MP, Thomas JP, Palmer J, Shimozaki SS (2000) A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. Perception & psychophysics 62(3):425–451

Eckstein MP, Koehler K, Welbourne LE, Akbas E (2017) Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes. Current Biology 27(18):2827–2832.e3, DOI 10.1016/j.cub.2017.07.068

Eliasmith C, Anderson CH (2003) Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems. MIT press

Eliasmith C, Stewart TC (????) Nengo and the Neural Engineering Framework: Connecting Cognitive Theory to Neuroscience p 3

Geisler WS (2003) Ideal observer analysis. The visual neurosciences 10(7):12–12

Geisler WS, Cormack LK (2011) Models of overt attention. Oxford handbook of eye movements pp 439–454

Gosmann J, Voelker AR, Eliasmith C (????) A Spiking Independent Accumulator Model for Winner-Take-All Computation p 6

Grossberg S, Mingolla E, Ross WD (????) A Neural Theory of Attentive Visual Search: Interactions of Boundary, Surface, Spatial, and Object Representations p 20

Hommel B, Chapman CS, Cisek P, Neyedli HF, Song JH, Welsh TN (2019) No one knows what attention is. Attention, Perception, & Psychophysics 81(7):2288–2303, DOI 10.3758/s13414-019-01846-w

Hulleman J, Olivers CNL (2017) The impending demise of the item in visual search. Behavioral and Brain Sciences 40:e132, DOI 10.1017/S0140525X15002794

Ionescu RT, Alexe B, Leordeanu M, Popescu M, Papadopoulos DP, Ferrari V (2016) How Hard Can It Be? Estimating the Difficulty of Visual Search in an Image. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, pp 2157–2166, DOI 10.1109/CVPR.2016.237

Jozwik KM, Kriegeskorte N, Storrs KR, Mur M (2017) Deep Convolutional Neural Networks Outperform Feature-Based But Not Categorical Models in Explaining Object Similarity Judgments. Frontiers in Psychology 8:1726, DOI 10.3389/fpsyg.2017.01726

Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ (2019) Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nature Neuroscience 22(6):974–983, DOI 10.1038/s41593-019-0392-5

Katti H, Peelen MV, Arun SP (2017) How do targets, nontargets, and scene context influence real-world object detection? Attention, Perception, & Psychophysics 79(7):2021–2036, DOI 10.3758/s13414-017-1359-9

Kell AJ, McDermott JH (2019) Deep neural network models of sensory systems: Windows onto the role of task constraints. Current opinion in neurobiology 55:121–132

Kietzmann TC, Spoerer CJ, Sörensen LKA, Cichy RM, Hauk O, Kriegeskorte N (2019) Recurrence is required to capture the representational dynamics of the human visual system. Proceedings of the National Academy of Sciences 116(43):21854–21863, DOI 10.1073/pnas.1905544116

Kornblith S, Shlens J, Le QV (2019) Do better imagenet models transfer better? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2661–2671

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp 1097–1105

Lindsay GW (2020) Attention in Psychology, Neuroscience, and Machine Learning. Frontiers in Computational Neuroscience 14:29, DOI 10.3389/fncom.2020.00029

Ma WJ, Navalpakkam V, Beck JM, van den Berg R, Pouget A (2011) Behavior and neural basis of near-optimal visual search. Nature Neuroscience 14(6):783–790, DOI 10.1038/nn.2814

Majaj NJ, Hong H, Solomon EA, DiCarlo JJ (2015) Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. Journal of Neuroscience 35(39):13402–13418, DOI 10.1523/JNEUROSCI.5181-14.2015

Moran R, Zehetleitner M, Muller HJ, Usher M (2013) Competitive guided search: Meeting the challenge of benchmark RT distributions. Journal of Vision 13(8):24–24, DOI 10.1167/13.8.24

Nakayama K, Martini P (2011) Situating visual search. Vision Research 51(13):1526–1537, DOI 10.1016/j.visres.2010.09.003

Narbutas V, Lin YS, Kristan M, Heinke D (2017) Serial versus parallel search: A model comparison approach based on reaction time distributions. Visual Cognition 25(1-3):306–325, DOI 10.1080/13506285.2017.1352055

Nayebi A, Bear D, Kubilius J, Kar K, Ganguli S, Sussillo D, DiCarlo JJ, Yamins DL (????) Task-Driven Convolutional Recurrent Models of the Visual System p 12

Nicholson D, Prinz A (2019) Convolutional neural networks performing a visual search task show attention-like limits on accuracy when trained to generalize across multiple search stimuli. In: 2019 Conference on Cognitive Computational Neuroscience, Cognitive Computational Neuroscience, Berlin, Germany, DOI 10.32470/CCN.2019.1432-0

Palmer J, Verghese P, Pavel M (2000) The psychophysics of visual search. Vision research 40(10-12):1227–1268

Peelen MV, Kastner S (2014) Attention in the real world: Toward understanding its neural basis. Trends in cognitive sciences 18(5):242–250

Poder E (2017) Capacity limitations of visual search in deep convolutional neural network

Rasmussen D (2019) NengoDL: Combining deep learning and neuromorphic modelling methods. arXiv:180511144 [cs] 1805.11144

Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, Clopath C, Costa RP, de Berker A, Ganguli S (2019) A deep learning framework for neuroscience. Nature neuroscience 22(11):1761–1770

Saxe A, Nelli S, Summerfield C (????) If deep learning is the answer, then what is the question? p 26

Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, Kar K, Bashivan P, Prescott-Roy J, Schmidt K (2018) Brain-Score: Which artificial neural network for object recognition is most brain-like? BioRxiv p 407007

Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:14091556 [cs] 1409.1556

Spoerer CJ, McClure P, Kriegeskorte N (2017) Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition. Frontiers in Psychology 8:1551, DOI 10.3389/fpsyg.2017.01551

Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cognitive psychology 12(1):97–136

Wolfe JM (1994) Guided search 2.0 a revised model of visual search. Psychonomic bulletin & review 1(2):202–238

Wolfe JM (1998) Visual search. In: Attention, Psychology Press/Erlbaum (UK) Taylor & Francis, Hove, England, pp 13–73

Wolfe JM, Horowitz TS (2017) Five factors that guide attention in visual search. Nature Human Behaviour 1(3):1–8, DOI 10.1038/s41562-017-0058

Wolfe JM, Cave KR, Franzel SL (1989) Guided search: An alternative to the feature integration model for visual search. Journal of Experimental Psychology: Human perception and performance 15(3):419

Wolfe JM, Palmer EM, Horowitz TS (2010) Reaction time distributions constrain models of visual search. Vision research 50(14):1304–1311

Wolfe JM, Alvarez GA, Rosenholtz R, Kuzmova YI, Sherman AM (2011) Visual search for arbitrary objects in real scenes. Attention, Perception, & Psychophysics 73(6):1650–1671, DOI 10.3758/s13414-011-0153-3

Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences 111(23):8619–8624

Yamins DLK, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience 19(3):356–365, DOI 10.1038/nn.4244

Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems, pp 3320–3328