Convolutional neural networks performing a visual search task show attention-like limits on accuracy when trained to generalize across multiple search stimuli

David A. Nicholson (dnicho4@emory.edu)

Astrid A. Prinz (astrid.prinz@emory.edu)
Emory University Department Of Biology, 1510 Clifton Road, Atlanta, GA 30322

Abstract

What limits our ability to find what we are looking for in the cluttered noisy world? To investigate this, cognitive scientists have long used visual search. In spite of hundreds of studies, it remains unclear how to relate effects found using the discrete item display search task to computations in the visual system. A separate thread of research has studied the visual system of humans and other primates using convolutional neural networks (CNNs) as models. Multiple lines of evidence suggest that training CNNs to perform tasks such as image classification causes them to learn representations similar to those used by the visual system. These studies raise the question of whether CNNs that have learned such representations behave similarly to humans performing other vision-based tasks. Here we address this by measuring the behavior of CNNs trained for image classification while they perform the discrete item display search task. We first show how a fine-tuning approach often used to adapt pre-trained CNNs to new tasks can produce models that show human-like limitations on this task. However we then demonstrate that we can greatly reduce these effects by changing training,without changing the learned representations. Lastly we show that accuracy is not impaired when single networks are trained to discriminate multiple types of visual search stimuli. Based on these findings, we suggest that CNNs are not necessarily subject to the same limitations as the primate visual system.

Keywords: attention;visual search

Introduction

What limits our ability to find what we are looking for in the cluttered noisy world we see around us? One of the principle tasks that has been used to investigate this question is a visual search task (Figure 1) (Wolfe, 1998a) we will refer to as the *discrete item display search* task. Most studies using this task experimentally manipulate factors such as features of the targets and distractors in order to identify those factors that limit visual search (Eckstein, 2011; Wolfe & Horowitz, 2017), known as *capacity limitations*. There are essentially two models of capacity limitations: *attention-limited* and *noise-limited* models (E. M. Palmer et al., 2011). Briefly, we review these and describe how they depend in part on the way the visual search task is performed. Both models depend on what are known as *set-size effects* seen when using visual search stimuli, depicted schematically in 1**b**). Attention-limited models

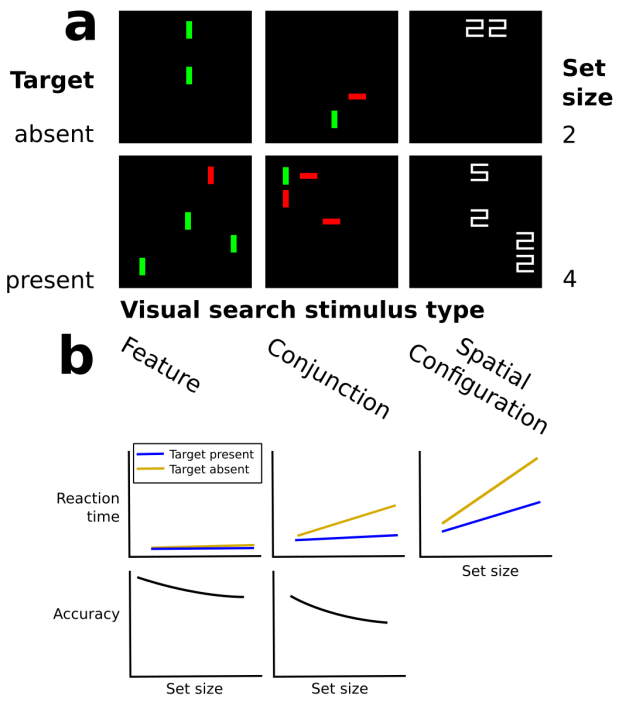

Figure 1: **The discrete item display search task**. On each trial, subjects view a display (example images, **a**) of discrete items on a flat background. Subjects respond whether the target is present or absent (condition varies across rows of **a**). Another condition that varies across trials is the set size, that is, the total number of targets and distractors (varies across rows in **a**). Many studies focus on set size effects. Typical effects are an increase in reaction time or a decrease in accuracy as set size increases (depicted schematically in **b**, redrawn from (Wolfe et al., 2010) and (Eckstein, 1998)). Effects vary based on the features that distinguish targets from distractors (shown in columns). Details in introduction. Accuracy for spatial configuration-type stimuli not shown in **b** because this has been less studied (but see (E. M. Palmer et al., 2011)).

posit two-stage theories of visual processing, a first *preattentive* stage which can process single features such as color or orientation in parallel, and a second *attentive* stage which requires a serial computation to binds features, resulting in a bottleneck (Treisman & Gelade, 1980; Wolfe, Cave, & Franzel,

1989; Wolfe, 1994). Evidence for this theory came from visual search experiments where subjects were shown stimuli until responding, and their reaction time was measured. Plotting reaction time as a function of set size revealed lines whose slope were near zero (1**b**, top row, left plot) when a single feature distinguished the target from distractors (1, **a**, left column). The slope of reaction time versus set size increased though (1**b**, top row, middle plot), when distinguishing the target from distractors required finding a conjunction of features (1**a**, middle column). Slopes became even steeper (1**b**, top row, right plot) when targets and distractors shared features and could only be distinguished by their spatial configuration (1**a**, right column). This sort of set size effect has been replicated hundreds of times, although it remains unclear why some search stimuli are processed effectively in parallel while others face a bottleneck (Wolfe, 1998b; Wolfe & Horowitz, 2017). The second family of capacity limitations conceives of visual search as completely parallel but noisy. These models arose in large part in reaction to the way the visual search task was carried out when investigating serial mechanisms. Early studies that measured reaction times left several factors uncontrolled, such as target-distractor similarity (Duncan & Humphreys, 1989), drops in acuity outside the fovea, eye movements, and effects resulting from visual crowding (Eckstein, 1998). Hence researchers designed versions of the visual search task that controlled for such factors. Crucially, they showed subjects the stimulus only briefly, to prevent eye movements, and measured accuracy instead of reaction time (shown schematically in 1**b**, bottom row). Computational models of parallel mechanisms, based on signal detection theory, successfully explained results from feature and conjunction search stimuli (1**a**, left and middle columns) (J. Palmer, Verghese, & Pavel, 2000; Eckstein, 2011).

In spite of nearly half a century of studies based on the discrete item display search task, this core question remains unresolved: to what extent can limitations be attributed to an attention-like computation, e.g. binding features into items, and to what extent can those limitations be attributed to other computations, e.g. a decision-making process subject to noisy internal representations? To foreshadow our approach, we suggest another way of posing this question: if some algorithm could produce a statistical model capable of learning from data to classify visual search stimuli as target present or absent with high accuracy, would that accuracy still be subject to some ceiling, due simply to the constraints of the task? This framing bears some similarity to a well-established framework in vision research known as ideal observer analysis (Geisler, 2003). However, few ideal observer models have taken the form of "pixel-in, behavior-out", and therefore do not generalize to many different types of visual search stimuli, so their predictive power is limited (Geisler & Cormack, 2011).

Both attention-limited and noise-limited models are highly abstracted models of the visual system, in which low-level features pass through a hierarchy until reaching a final stage consisting of a simple decision rule. Similarly, the architecture of convolutional neural networks (CNNs) now routinely used for computer vision tasks represents a highly abstracted view of the visual system. Many researchers have drawn parallels between the architecture of CNNs and the architecture of the visual system in the brain (Kriegeskorte, 2015), in particular in humans and other primates where this system has been most thoroughly studied. Like CNNs, the visual system has a hierarchical structure, and is thought to function in part by performing transformations at each level of this hierarchy so that high-dimensional, low-level features are mapped into low-dimensional abstract representations. Several recent studies have found that that, when optimized to perform tasks such as image classification, CNNs learn representations similar to those observed in the visual system (D. L. K. Yamins & DiCarlo, 2016; D. L. Yamins et al., 2014). These studies raise the question of whether CNNs that have learned such representations behave similarly to humans performing other vision-based tasks. Here we address this by measuring the behavior of CNNs trained for image classification while they perform the discrete item display search task. While there have been previous studies of neural networks performing visual search tasks, we are aware of only one study (Poder, 2017) that employed the sort of CNN architectures used in studies of the visual system referenced above. We replicate the methods from that study, and extend that author's results. In the interest of replicability, we have made the code and summary results [1] available, and will release the raw data upon publication [2].

Results

As referenced above, previous work suggests that optimizing CNNs to perform image classification causes them to acquire representations which resemble those that can be identified in the brain. We first tested whether the CNN architecture AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) would be subject to human-like limits on the discrete item display search task, when utilizing representations learned by optimizing the weights for image classification on the ImageNet dataset (Deng et al., 2009). To do so, we used a *fine-tuning* approach, in which we used pre-trained weights in earlier convolutional layers of AlexNet, but randomly initialized weights in the later fully-connected layers (Yosinski, Clune, Bengio, & Lipson, 2014). More specifically, we replicated the training method described in (Poder, 2017), where we trained each network that we tested with 6400 samples of *one* of the visual search stimuli (e.g. feature search), and then measured accuracy on a separate test set of 800 samples. Stimuli were generated with a small Python package [3] which produced images the same size as the images used to train on ImageNet, and were pre-processed in the same way as those images, except that no re-sizing was done. This training method produced AlexNet models whose accuracy showed set-size

[1] https://github.com/NickleDave/visual-search-nets

[2] at https://figshare.com/articles/visual-search-nets/7688840

[3] https://github.com/NickleDave/searchstims

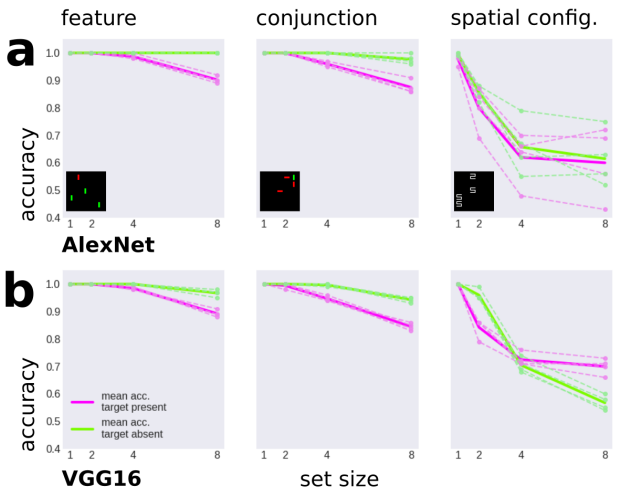

Figure 2: **Human-like limits on accuracy of CNNs performing the discrete item display search task** As described in the text, when using weights pre-trained on ImageNet in convolutional layers, and fine-tuning weights in fully-connected layers so AlexNet could perform the task, this CNN showed decreases in accuracy as set size increased (**a**). This decrease was smallest for feature search stimuli (left plot), intermediate for conjunction stimuli (middle plot), and largest for spatial configuration stimuli (right plot). The same approach produced similar results with the VGG16 architecture (**b**).

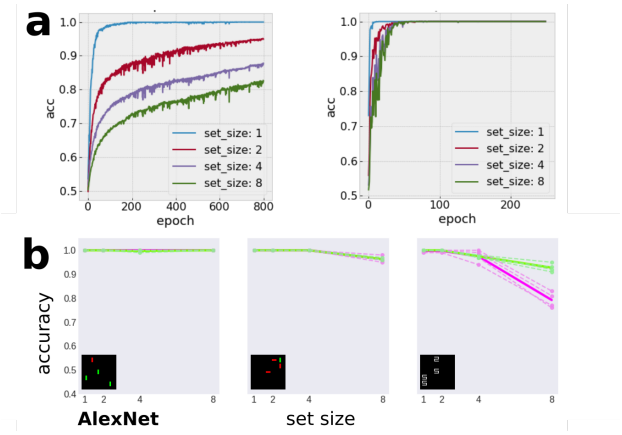

Figure 3: **Changing training greatly reduces set size effects.** Training histories showed that the accuracy of models trained with the fine tuning approach did not converge on some asymptotic value, and varied depending on the set size of the search stimuli (**a**, left plot). Increasing the learning rate and including more examples of stimuli with larger set sizes greatly sped up convergence (**a**, right plot). AlexNet models trained with this higher learning rate and larger training set showed reduced set size effects (**b**).

dependence similar to human subjects (Figure 2**a**). For the three types of visual search stimuli that we used, there was a set size effect, where accuracy decreased as the set size increased. This effect was smallest for feature search (left column), slightly larger for conjunction search (middle column), and largest for spatial configuration search. To test whether this effect was unique to AlexNet, we also used the same approach with the VGG16 architecture, and produced similar results (Figure 2**b**). In addition to finding set size effects that were qualitatively similar to those seen in human subjects, we also noted that accuracy was always higher for the "target absent" condition. We do not find reports of similar differences between target present and target absent conditions in human subjects.

Because accuracy of CNNs depends in part on training, it could be the case that the results just described are an artifact of how we trained the networks. To gain insight into how our results depended on training, we plotted training histories where we measured accuracy on the training set at each epoch *separately for each set size in the visual search stimuli*. These plots revealed (1) that accuracy had not yet approached some asymptotic value by the end of training, and (2) that there was an inverse relationship between the set size of a visual search stimulus and the rate that its accuracy increased, e.g. accuracy on set size 1 reached its highest value within a few epochs, while accuracy on set size 8 never converged on an asymptotic value (figure 3**a**). Because these plots showed accuracy had not converged when we stopped training, we used random search to find an optimal learning rate, a key hyperparameter. We did find we were able to improve accuracy and decrease training time by using a typical learning rate on the fully-connected layers, while simply freezing the pre-trained weights in earlier convolutional layers. Based on the the observation that different set sizes converged at different rates, we also augmented the number of samples for larger set sizes. It may seem counterintuitive to "unbalance" the dataset this way, until one considers that there are many more combinations of displays of set size 8 then of set size 1. For example, if items are located on a 5-by-5 grid, then for set size 1 there are 25-pick-1 combinations, i.e. 25, while for set size 8 there are 25-pick-8 combinations (approximately 100k). (Note that the library we used to generate images ensures that there are no repeats; for set size 1, jitter is added to produce more than 25 possible displays; this is typically done in experiments with human subjects, and acts as a form of data augmentation for neural networks.) We simply multiplied the number of samples by the set size, since scaling by the number of possible combinations would have produced prohibitively large datasets. After making these changes to the learning rate and the statistics of the dataset, we saw that the set size effects were greatly reduced 3**b**. This indicates that these effects are due in least at part to how we trained the networks.

Because neural networks, including CNNs, act as function approximators, it could also be the case that they were able

to perform the discrete item display search task with relatively high accuracy simply by learning an exclusive-or function for the single search stimulus that we trained them to classify. A more rigorous test would be to train single networks to classify multiple stimuli, in the same way that humans do in discrete item display search experiments. As a final test, we trained two instances of Alexnet on very large datasets containing all three stimuli used in this study 4. We found similar accuracy as shown in 3, even when single networks were challenged to perform this task with multiple stimuli.

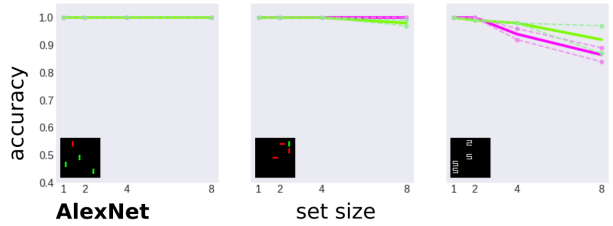

Figure 4: **Training single networks on multiple stimuli does not impair accuracy** Two instances of AlexNet trained on datasets containing all three types of visual search stimuli used in this study still attained high accuracy.

Discussion

We tested whether CNNs using representations learned from image classification tasks would be subject to human-like capacity limitations when performing the discrete item display search task. While we did not find that CNNs can always perform the discrete item display search task with perfect accuracy, we did show that these models are not necessarily limited by the same factors as the primate visual system.

Acknowledgments

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. Thank you to Zsolt Kira and Yen-Chang Hsu for feedback on results, and for suggesting changes to training methods. Thank you also to Constantine Dovrolis and Sarah Pallas for feedback on earlier versions of this work.

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee.

Duncan, J., & Humphreys, G. W. (1989). Visual Search and Stimulus Similarity. , *96*(3), 433-458.

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science*, *9*(2), 111-118.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of vision*, *11*(5), 14-14.

Geisler, W. S. (2003). Ideal observer analysis. *The visual neurosciences*, *10*(7), 12-12.

Geisler, W. S., & Cormack, L. K. (2011). Models of overt attention. *Oxford handbook of eye movements*, 439-454.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*, 417-446.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (p. 1097-1105).

Palmer, E. M., Fencsik, D. E., Flusberg, S. J., Horowitz, T. S., & Wolfe, J. M. (2011). Signal detection evidence for limited capacity in visual search. *Attention, Perception, & Psychophysics*, *73*(8), 2413-2424.

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

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

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97-136.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, *1*(2), 202-238.

Wolfe, J. M. (1998a). Visual search. In *Attention* (p. 13-73). Hove, England: Psychology Press/Erlbaum (UK) Taylor & Francis.

Wolfe, J. M. (1998b). What can 1 million trials tell us about visual search? *Psychological Science*, *9*(1), 33-39.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (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, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*(3), 1-8. doi: 10.1038/s41562-017-0058

Wolfe, J. M., Palmer, E. M., & Horowitz, T. S. (2010). Reaction time distributions constrain models of visual search. *Vision research*, *50*(14), 1304-1311.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yamins, D. L. K., & DiCarlo, J. J. (2016, March). 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).