Capacity Limitations of Visual Search in Deep Convolutional Neural Networks

Endel Pöder  
endel.poder@ut.ee  
Institute of Psychology, University of Tartu, 50409 Tartu, Estonia

Deep convolutional neural networks (CNN) follow roughly the architecture of biological visual systems and have shown a performance comparable to human observers in object classification tasks. In this study, three deep neural networks pretrained for image classification were tested in visual search for simple features and for feature configurations. The results reveal a qualitative difference from human performance. It appears that there is no clear difference between searches for simple features that pop out in experiments with humans and for feature configurations that exhibit strict capacity limitations in human vision. Both types of stimuli reveal comparable capacity limitations in the neural networks tested here.

1 Introduction

It is well known that human observers have certain limitations on simultaneous processing of multiple visual stimuli (Estes & Taylor, 1964; Bergen & Julesz, 1983). Visual search experiments have revealed simple features (luminance, color, size, orientation) that can be detected in parallel across the visual field, independent of the number of objects (Wolfe, 1998). Detection of combinations of simple features is more difficult and may need serial processing (Treisman & Gelade, 1980; Wolfe, Cave, & Franzel, 1989).

The majority of early visual search studies used reaction time as the measure of performance. However, qualitatively comparable results can be found using a brief presentation of stimuli and accuracy as a measure of performance. Data from these experiments can be naturally analyzed in terms of signal detection theory (SDT; Kinchla, 1974; Palmer, Verghese, & Pavel, 2000). SDT assumes noisy representation of visual items and (nearly) optimal decision making. Different behavior of simple and complex features is still important. Searching for a simple feature among homogeneous distractors fits well to an SDT model that assumes independent encoding of visual objects and ideal integration of noisy signals (Shaw, 1984; Palmer, Ames, & Lindsey, 1993; Palmer, 1994). Searching for configurations of simple features has strictly limited capacity and exhibits set size effects consistent with dividing fixed processing resources, or serial scanning (Shaw, 1984; Pöder,
Figure 1: Theoretical signal detection models of visual search (top). The unlimited capacity model (top left) assumes that the precision of encoding of individual items is independent of the number of items, and the fixed capacity model (top right) assumes that the variance of individual representations is proportional to the number of items. Curves with different colors correspond to different levels of target-distractor discriminability. (Bottom) An example of human visual search data: simple feature search (blue) and configuration search (red). $b$ is a measure of encoding capacity limitations.

1999; Palmer, Fencsik, Flusberg, Horowitz, & Wolfe, 2011). Figure 1 depicts predictions of two classic SDT-based search models and an example of respective human search data. Although both models predict a deterioration of performance with increasing set size, as well as with decreasing target-distractor discriminability, the set size effect is much stronger in the limited capacity model. These models can be unified using a quantitative measure of set size effect. The unified model assumes that encoding variance is a power function of set size, $\nu \propto n^b$, where $b = 0$ for an unlimited capacity and $b = 1$ for a fixed capacity (sample size) model.
According to a widely accepted view, spatial attention plays an important role in the perception of complex objects (Treisman & Gelade, 1980; Cheal, Lyon, & Hubbard, 1991; Wolfe & Bennett, 1996). It is believed that spatial attention selects relevant visual signals at relatively low levels and in retinotopic coordinates and thus simplifies processing at higher levels (Broadbent, 1958; Neisser, 1967). However, there are different opinions (Deutsch & Deutsch, 1963; Allport, Tipper & Chmiel, 1985). In recent studies, Rosenholtz (2017; Rosenholtz, Huang, & Ehinger, 2012) has argued that spatial gating is not necessary in visual processing, and apparent capacity limitations may reflect the complexity of a decision boundary in some high-level multidimensional space where a representation of the whole visual field is encoded. This view resembles information processing in current artificial neural networks. Therefore, experimenting with these networks may help to test theories of biological vision as well.

About ten years ago, artificial neural networks reached the level of human performance in demanding visual object classification tasks (Krizhevsky, Sutskever, & Hinton, 2012; Ciresan, Meier, & Schmidhuber, 2012; Simonyan & Zisserman, 2014; Szegedy et al., 2015). These networks are hierarchical feature combiners following roughly the architecture of biological visual systems and trained on millions of labeled natural images. Many studies have reported on functional similarities between deep neural networks and visual systems of humans or monkeys (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Kubilius, Bracci, & Op de Beeck, 2016). However, several interesting differences have been reported as well (Nguyen, Yosinski, & Clune, 2015; Geirhos et al., 2018; Kim et al., 2018; Lonnqvist, Clarke, & Chakravarthi, 2020). And a few studies in visual search have reported on both similarities (Gupta, Zhang, Wu, Wolfe, & Kreiman, 2021) and differences (Eckstein, Koehler, Welbourne, & Akbas, 2017) between neural networks and human observers.

Zhang et al. (2018) used a pretrained CNN model (VGG 16) as a basic component in their visual search model. They extracted features from the last convolutional layer, calculated local feature similarity to the target features, and created an “attention map” to guide eye movements in the input display. The model predicted human reaction times of searching for targets in natural and cartoon images reasonably well. These results suggest that complex features learned for object classification can support a visual search task as well. However, the search for simple visual features was not examined in this study.

Up to now, a classic topic in human vision—the extent of capacity limitations in processing of simple and complex features—has not been studied in neural networks. Usually neural networks do not contain any mechanisms of spatial attention. Assuming the important role of attention in humans’ visual searches, the search with neural networks might be very different from search with humans. However, SDT-based search models do not assume mechanisms of attention and could well be implemented by standard
neural networks. Similar capacity limitations of human observers and pre-trained neural networks might indicate that these are determined by the complexity of visual tasks rather than specific mechanisms like attention. Certainly a comparison of visual search regularities of artificial neural networks with findings from human observers may reveal interesting aspects of visual processing.

In this study, simple search experiments were run using three pretrained convolutional neural networks in place of a human observer. Two types of stimuli, corresponding to simple feature and feature configuration search, were used, and the results were compared with our knowledge of human performance.

Several studies have revealed that neural networks with random weights can be surprisingly effective in several tasks (Jarrett, Kavukcuoglu, Ranzato, & LeCun, 2009; Saxe et al., 2011). Therefore, finding certain regularities of visual search in neural networks pretrained for image classification does not necessarily mean that these regularities are caused by the representations acquired during the pretraining. To clarify the role of pretraining, I replicated my search experiments using the same networks with initial random weights instead of these trained for image classification.

2 Methods

In the experiments presented in this letter, the pretrained neural networks AlexNet, GoogLeNet, and ResNet18 provided with the Matlab Deep Learning Toolbox, and a transfer learning procedure, were used. The final three layers (fully connected, softmax, and output) originally configured for the classification of 1000 natural image categories of the ImageNet were replaced with equivalent layers for the classification into two categories: target present and target absent. Only the fully connected one among the new layers contained trainable weights and biases. These parameters were adjusted during the training with visual search stimuli.

Search stimuli were generated in Matlab. The size of stimuli was 227 × 227 or 224 × 224 pixels × 3 color planes. Each image contained $n$ ($n = 1, 2, 4, \text{or } 8$) simple items (squares, lines, rectangles, rotated Ts) depicted on a dark background. To minimize possible spatial interactions, the minimal center-to-center distance between the items was set to be at least 48 pixels. Also, the items were not placed within 28-pixel edges of the image. Otherwise, the items were located randomly. The images of the target present category contained one target item and $n - 1$ distractor items, and the images of the target absent category contained only $n$ distractors. The target item was fixed during an experiment.

In this study, six search experiments with different visual features were run (examples of stimuli are given in Figure 2). There were four “simple” tasks, with targets of either different luminance, color, length, or orientation,
Figure 2: Examples of visual search stimuli used in this study. All examples depict stimuli with the target present. Set size (number of objects in a display) was varied from 1 to 8.

and two “complex” tasks: rotated Ts, where the target differs from distractors by spatial configuration of two bars, and bicolor squares, where the target and distractors have the same colors in different order. For rotated Ts, four versions of the experiment were run, with each orientation of the letter selected as the target (the remaining three orientations were used as distractors, randomly selected).

The difficulty levels for search tasks were chosen to minimize both perfect and chance level performance. From four to six different difficulty levels were used for each search experiment. For simple features, difficulty was manipulated using target-distractor difference (e.g., difference in brightness, or bar length). For rotated Ts and bicolor squares, the size of items was varied to that purpose. Different difficulty levels were trained and tested in separate procedures. All other parameters of stimuli, not mentioned here, were held constant during the experiments.

It is believed that human observers can be more efficient when a set size is fixed within blocks of trials because of reduced uncertainty. For a neural network, transfer learning with a fixed set size could be superior too because it allows finding the optimal weights for that condition, without interference from other set sizes. The results reported in this study are based on separate training of set sizes. However, my preliminary experiments (Pöder,
indicate that training with mixed set sizes could produce qualitatively similar results.

To meaningfully compare performance across different set sizes, difficulty levels, and tested networks, the transfer learning should approach asymptotic performance in all the conditions. In pilot experiments, I searched for hyperparameters that could support the efficient convergence to the nearly best performance across different conditions. In the final simulations, training sets of 6400 images were used, 3200 of target present, and 3200 of target absent category. The validation set consisted of 800 images. Training with stochastic gradient descent was used. Minibatch size was 200, and 400 epochs through a training set were run. The initial learning rates were 0.003, 0.03, or 0.3 for AlexNet, GoogLeNet, and ResNet18, respectively. Learning rate was then reduced by factor of 2 after every 80 epochs. The hyperparameters chosen are not too critical. Varying training set size, number of epochs, or learning rate by a factor of 2 could change final performance by only 1 or 2 percentage points.

In standard transfer learning, new input stimuli are repetitively passed through the whole network, but only the weights in new layers are updated. To reduce computing time, I separated the old and new parts of the network. I passed my new training set through the original pretrained network and saved the activations from the last task-independent layer. Then I used the saved activations as input features to train a small new network consisting of an input layer and three task-specific output layers. While this method is functionally equivalent to usual transfer learning, it can greatly reduce training time.

Two simple models were used to measure the effects of set size on proportion correct. The first one is based on an observation that $d'$ prime versus set size relationship is approximately linear in log-log graphs, and the curves for different target-distractor discriminability are roughly parallel. To measure set size effect, I estimated the log-log slope of these curves. This model does not make specific predictions for different stimuli, but assuming human-like regularities, the slopes should be much larger for feature configuration compared to simple feature search.

The second is an SDT-based search model with (possibly) limited capacity encoding and ideal decision rules (Palmer et al., 2000; Mazyar, Van den Berg, & Ma, 2012; Pöder, 2017). This model measures the effect of set size on encoding precision (noise variance). This measure is 0 for unlimited capacity (independent processing of items) and 1 for a fixed capacity (encoding noise variance proportional to set size). The two models have equal numbers of free parameters and are easily comparable.

To test the role of ImageNet-based pretraining, I replicated my search experiments using the same networks with initial random weights instead of these trained for image classification. Otherwise, the transfer learning and testing procedure was identical.
3 Results

Typical training curves are shown in Figure 3. Apparently the curves of different set sizes (as well as different difficulty levels) converge to different asymptotic levels. With bigger set sizes, learning becomes noisier as well. More detailed training curves, including accuracy and cross-entropy loss for training (minibatches) and validation sets, are given in supplementary Figure S1.

Examples of elementary data sets from experiments with different features and respective model fits are depicted in Figure 4. Other data sets were qualitatively similar (they are given in supplementary figures S2 to S4). The model fits were not perfect, but both models capture the main regularities of data reasonably well. The goodness-of-fit statistics for the two models are given in Table 1. It appears that the data favor SDT-based limited capacity model. While 8 elementary data sets (out from 27) showed significant ($p < 0.01$) deviation from the SDT model, 15 deviated from the constant slope model. The average difference in the $G$ statistic was 6.8, which corresponds to an average ratio of likelihoods of about 30, in favor of SDT model.

The estimated measures of set size effects are depicted in Figure 5 (for rotated Ts, average results over four target orientations are given). Most interesting, there is no clear difference between classic simple features and feature configuration (rotated T, bicolor square) search. In both conditions, neural networks exhibit moderate to strong capacity limitations. Quantitatively, the average set size effects were slightly stronger for feature configuration compared to simple feature search: 0.77 versus 0.71 in terms of log-log slope and 0.97 versus 0.85 in terms of capacity limitations. The last difference was statistically significant, with $p < 0.05$. 

Figure 3: An example of training curves for different set sizes, here training AlexNet to search for a bright rectangle among dimmer ones. Target-distractor difference is fixed.
Figure 4: Examples of elementary data sets. The proportions correct as dependent on set size and target-distractor discriminability for different search experiments. Symbols depict experimental data, and lines are fits of the SDT model. Different colors correspond to different target-distractor difference.

Still, both values are rather close to the fixed capacity prediction \((b = 1)\) and far from the unlimited capacity \((b = 0)\) usually observed with humans in simple feature searches. The data also show some variance across simple features: brightness and length exhibited somewhat smaller set size effects compared to color and orientation. There were no significant differences in the networks studied.
Table 1: Model Fits (Likelihood Ratio Statistic $G$) of the Results from Search Experiments.

|                        | Fixed Log-Log Slope | SDT Model |
|------------------------|---------------------|-----------|
|                        | AlexNet | GoogLeNet | ResNet18 | AlexNet | GoogLeNet | ResNet18 |
| Brightness             | 34.7*   | 13.5      | 37.2*    | 16.9    | 18.8      | 32.2*    |
| Length                 | 22.2    | 37.2*     | 60.9*    | 12.4    | 32.2      | 30.2*    |
| Color                  | 82.5*   | 17.1      | 20.3     | 52.4*   | 11.4      | 14.2     |
| Orientation            | 51.7*   | 40.8*     | 98.8*    | 30.6*   | 25.8      | 96.7*    |
| Rotated T1             | 20.6    | 15        | 31.8*    | 21.8    | 22.6      | 16.2     |
| Rotated T2             | 14.8    | 32.6*     | 51.9*    | 12.7    | 23.2      | 29.4*    |
| Rotated T3             | 8.7     | 15.1      | 33.3*    | 11.5    | 22.7      | 41.2*    |
| Rotated T4             | 13.3    | 14.4      | 34.3*    | 13      | 24.7      | 26.3     |
| Bicolor squares        | 35.7*   | 22        | 29.9*    | 25.5    | 27.5      | 30.4*    |

Note: Significant differences (with $p < 0.01$) between data and models are indicated.

Figure 5: Set size effects as $d$ prime versus set size log-log slopes (top) and SDT-based capacity limitation $b$ (bottom) for different search experiments and for different neural networks.
The data from networks not trained for image classification are given in supplementary Figures S2 to S4, together with corresponding graphs from pretrained networks. The results were different for different networks and for different search tasks. Most frequently, performance with random weights was considerably lower as compared to the pretrained networks. This pattern was observed for length, color, and rotated T search for all three networks. However, the drop was most remarkable for AlexNet. Still, in some conditions, random weights seem to work even better than those adapted for image classification. In the brightness search, performance with random weights was much better compared to the pretrained networks for GoogLeNet and ResNet18. These networks also exhibited some superiority of random weights in search for a bicolor target.

4 Discussion

In this study, simple visual search experiments were run on pretrained deep convolutional neural networks. The results show that representations created by training CNNs for natural image classification can also support a moderately efficient visual search. However, closer examination reveals important differences from biological vision. While studies with human observers have found big differences between searches for simple features and for feature configurations, virtually no difference was found with artificial neural networks. Both types of stimuli revealed moderate to strong capacity limitations in the studied neural networks.

The observed measure of capacity limitation was close to the prediction of a fixed capacity model \( b = 1 \) that has been used to account for feature configuration search in human experiments. When searching for simple visual features, human observers usually exhibit very small set size effects consistent with the unlimited capacity SDT model (e.g., Shaw, 1984; Palmer, 1994; Põder, 1999). This corresponds to \( b = 0 \) in terms of our generalized model. The simulation results with neural networks are very different from these findings.

There is no question that deep convolutional networks can learn to accomplish these simple tasks much better when allowed to adapt weights in the lower layers. Nicholson and Prinz (2021) have shown that AlexNet and VGG 16 are able to reach nearly perfect performance and zero set size effect in similar search tasks, when all the layers were trainable. However, the purpose of this study was to examine how well the learned image transformations necessary for object classification support visual search. The answer is “apparently not very well,” at least for simple features used in this study.

The results have some resemblance to my recent findings (Põder, 2021) on a CNN-based attention guidance model (Zhang et al., 2018). While complex visual features from the last retinotopic layer of the pretrained VGG 16 neural network accounted reasonably well for the search efficiency with
natural objects as stimuli, they remarkably underestimated performance in search of simple features and feature conjunctions. Both studies may indicate an inefficiency of encoding visual saliency of simple features in standard image classification CNNs.

The results with untrained networks revealed that even random weights can carry useful information to the output layers of a complex network. Obviously, the set of features computed by a network is a function of both weights and architecture. It seems that in more sophisticated architectures (GoogLeNet, ResNet18), useful transformations emerge more likely even without training. In a simple architecture like AlexNet, these transformations must be encoded in weights and can be acquired through training. As a result of training in the same task, the behavior of simple and complex networks becomes more similar.

While both training and architecture affect performance, it remains to be determined which details in network design or training history could produce human-like pop-out of simple feature singletons.

Efficient detection of feature singletons in biological vision has been related to saliency calculation through lateral inhibition between similar features (Wolfe et al., 1989; Nothdurft et al., 1999; Li, 2002). Lateral inhibition layers have been tested in neural networks as well (Jarrett, Gallant, & Van Essen, 2009) but not used in popular CNNs (note that a local response normalization used in AlexNet has no interactions across spatial locations and cannot support spatial saliency computation). Still, these functions can be learned through training if necessary. Possibly these mechanisms are unimportant for object classification but could play a more important role in visual search.

Are there any theoretical ideas that could predict quantitatively the set size effects from this study? Some properties of linear networks may be relevant. When the same set of units is used for a distributed coding of several variables, then interference from irrelevant ones appears as a noise in decoded pattern, with the variance proportional to the number of irrelevant variables. This could predict an approximately square root decay of precision as dependent on a number of items. The square root decay is also consistent with a linear decision rule in multidimensional feature space instead of the ideal one. However, the observed $d$ prime versus set size slopes was significantly larger than 0.5 in this study and not necessarily determined by a single mechanism.

The goodness-of-fit statistics showed a preference for an SDT-based search model over simple power function. This model combines an optimal decision rule with possible encoding limitations. Set size effects measured in this study imply relatively strong encoding limitations for both simple and complex stimuli. These could be explained by interference between the distributed representations of visual items. However, the observed limitations could be caused by an inability to apply an optimal decision rule as well. My experiments (Pöder, 2021) with the Zhang et al. (2018) model seem
to favor the first option because combining the features extracted from the 
last convolution layer of CNN with a separate, nearly optimal decision rule 
revealed a similar inefficiency of simple feature search.

This study resembles a recent article by Jacob, Pramod, Katti, and Arun 
(2021). These authors followed a similar idea to compare known qualitative 
findings from human studies with behavior of neural networks pretrained 
for object classification. Instead of transfer learning, they used a simpler 
method of Euclidean distance between the output feature vectors. While 
several classic phenomena of human vision were observed in a pretrained 
CNN as well, others were not. Our study allows adding efficient feature-
based search to the last category.

5 Conclusion

Using a transfer learning procedure, CNNs pretrained for object classifica-
tion, can be easily adapted to classic visual search tasks. However, these net-
works behave differently from human observers in the same tasks. Whereas 
human observers can search for a simple visual feature among homoge-
neous distractors very efficiently, standard neural networks cannot. They 
exhibit comparable set size effects in both simple feature searches and 
complex searches for feature configurations. These findings indicate that 
training for object classification does not build representations required for 
efficient search. Human-like search apparently requires learning special fea-
tures at least in several layers of a network and may be facilitated by some 
built-in properties of network architecture.

Acknowledgments

This study was supported by the Estonian Research Council (grant 
PUT663.)

References

Allport, D. A., Tipper, S. P., & Chmiel, N. R. J. (1985). Perceptual integration and 
post-categorical filtering. In M. I. Posner & O. S. M. Marin (Eds.), Attention and 
performance XI (pp. 107–132). Hillsdale, NJ: Erlbaum.
Bergen, J. R., & Julesz, B. (1983). Parallel versus serial processing in rapid pattern 
discrimination. Nature, 303, 696–698. 10.1038/303696a0, PubMed: 6855915
Broadbent, D. E. (1958). Perception and communication. London: Pergamon Press.
Cheal, M. L., Lyon, D. R., & Hubbard, D. C. (1991). Does attention have different 
effects on line orientation and line arrangement discrimination? Quarterly Journal 
of Experimental Psychology, 43A, 825–857. 10.1080/14640749108400959
Ciresan, D. C., Meier, U., Schmidhuber, J. (2012). Multi-column deep neural net-
works for image classification. In Proceedings of the IEEE Conference on Computer 
Vision and Pattern Recognition (pp. 3642–3649). Piscataway, NJ: IEEE.
Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review, 70*, 80–90. 10.1037/h0039515, PubMed: 14027390

Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology, 27*(18), 2827–2832. 10.1016/j.cub.2017.07.068, PubMed: 28889976

Estes, W. K., & Taylor, H. A. (1964). A detection method and probabilistic models for assessing information processing from brief visual displays. In *Proceedings of the National Academy of Sciences of the United States of America, 52*, 446–454. 10.1073/pnas.52.2.446

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalization in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems, 31*. Red Hook, NY: Curran.

Gupta, S. K., Zhang, M., Wu, C., Wolfe, J. M., & Kreiman, G. (2021). Visual search asymmetry: Deep nets and humans share similar inherent biases. arXiv:2106.02953v2.

Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications 12*, 1872. 10.1038/s41467-021-22078-3

Jarrett, K., Kavukcuoglu, K., Ranzato, M. A., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Proceedings of the IEEE 12th International Conference on Computer Vision* (pp. 2146–2153). Piscataway, NJ: IEEE.

Khaligh-Razavi, S-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Comput. Biol.*, 10(11), e1003915. 10.1371/journal.pcbi.1003915, PubMed: 25375136

Kim, J., Ricci, M., & Serre T. (2018). Not-So-CLEVR: Learning same-different relations strains feedforward neural networks. *Interface Focus 8*, 20180011. 10.1098/rsfs.2018.0011, PubMed: 29951191

Kinchla, R. A. (1974). Detecting target elements in multi-element arrays: A confusability model. *Perception and Psychophysics, 15*, 149–158. 10.3758/BF03205843

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 25* (pp. 1097–1105). Red Hook, NY: Curran.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLOS Comput. Biol.*, 12(4), e1004896. 10.1371/journal.pcbi.1004896, PubMed: 27124699

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*(1), 9–16. 10.1016/S1364-6613(00)01817-9, PubMed: 11849610

Lonnqvist, B., Clarke, A. D. F, & Chakravarthi, R. (2020). Crowding in humans is unlike that in convolutional neural networks. *Neural Networks, 126*, 262–274. 10.1016/j.neunet.2020.03.021, PubMed: 32272430

Mazyar, H., Van den Berg, R., & Ma, W. J. (2012). Does precision decrease with set size? *Journal of Vision, 12*(6).

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE*
Conference on Computer Vision and Pattern Recognition (pp. 427–436). Piscataway, NJ: IEEE.

Nicholson, D., & Prinz, A. (2021). Deep neural network models of object recognition exhibit human-like limitations when performing visual search tasks. bioRxiv. https://www.biorxiv.org/content/10.1101/2020.10.26.354258v2

Nothdurft, H. C., Gallant, J. L., & Van Essen, D. C. (1999). Response modulation by texture surround in primate area V1: Correlates of “popout” under anesthesia. Visual Neuroscience, 16, 15–34. 10.1017/S0952523899156189, PubMed: 10022475

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 andPsychophysics, 73, 2413–2424. 10.3758/s13414-011-0199-2

Palmer, J. (1994). Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks. Vision Research, 34, 1703–1721. 10.1016/0042-6989(94)0128-7, PubMed: 7941377

Palmer, J., Ames, C. T., & Lindsey, D. T. (1993). Measuring the effect of attention on simple visual search. Journal of Experimental Psychology: Human, Perception and Performance, 19, 108–130. 10.1037/0096-1523.19.1.108

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. Vision Research, 40, 1227–1268. 10.1016/S0042-6989(99)00244-8, PubMed: 10788638

Pöder, E. (1999). Search for feature and for relative position: Measurement of capacity limitations. Vision Research, 39, 1321–1327.

Pöder, E. (2017). Combining local and global limitations of visual search. Journal of Vision, 17(4), 10, 1–12.

Pöder, E. (2019). Visual search with deep convolutional neural network. Perception 48(25), 134.

Pöder, E. (2021). CNN-based search model underestimates attention guidance by simple visual features. arXiv:2103.15439.

Rosenholtz, R. (2017). Capacity limits and how the visual system copes with them. Journal of Imaging Science and Technology, 16, 8–23.

Rosenholtz, R., Huang, J., & Ehinger, K. A. (2012). Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. Frontiers in Psychology, 3(13), 1–15. 10.3389/fpsyg.2012.00013, PubMed: 22279440

Saxe, A., Koh, P. W., Chen, Z., Bhand, M., Suresh, B., & Ng, A. Y. (2011). On random weights and unsupervised feature learning. In Proceedings of the 28th International Conference on Machine Learning (pp. 1089–1096).

Shaw, M. L. (1984). Division of attention among spatial locations: A fundamental difference between detection of letters and detection of luminance increments. In H. Bouma & D. G. Bouwhais (Eds.), Attention and performance X (pp. 109–121). Hillsdale, NJ: Erlbaum.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1–9). Piscataway, NJ: IEEE.

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. Cognitive Psychology, 12, 97–136. 10.1016/0010-0285(80)90005-5, PubMed: 7351125
Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.), Attention (pp. 13–74). Hove, East Sussex, UK: Psychology Press.

Wolfe, J. M., & Bennett, S. C. (1996). Preattentive object files: Shapeless bundles of basic features. Vision Research, 37, 25–43. 10.1016/S0042-6989(96)00111-3

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration theory of attention. Journal of Experimental Psychology: Human Perception and Performance, 15, 419–433. 10.1037/0096-1523.15.3.419, PubMed: 2527952

Yamins, D. L. K., 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. PNAS, 111, 8619–8624. 10.1073/pnas.1403112111, PubMed: 24812127

Zhang, M., Feng, J., Ma, K. T., Lim, J. H., Zhao, Q., & Kreiman, G. (2018). Finding any Waldo with zero-shot invariant and efficient visual search. Nature Communications, 9(1), 3730. 10.1038/s41467-018-06217-x, PubMed: 30213937

Received November 9, 2021; accepted June 30, 2022.