PNAS COMMENTARY

How attention-like behavior emerges in an artificial neural network

Angus F. Chapman[a,1] and Rachel N. Denison[a]

Every time we look out for cars as we cross the street, search for a friend in a crowded restaurant, or listen for the announcement of our upcoming stop in a noisy subway car, we are using the cognitive process of attention. Attention allows us to prioritize relevant information while ignoring irrelevant information in the service of our current goals. An important feature of attention is its selectivity: Although we might wish we could fully process all the sensory information, we take in at every moment, attention selectively enhances some parts of the world at the expense of others (1, 2). Most research on attention has presumed the existence of dedicated neural mechanisms for attentional prioritization and selection. In a new study in PNAS, Srivastava et al. (3) instead investigate how attention-like behavior can emerge from an artificial neural network trained on a visual task.

The selective nature of covert visual spatial attention—the prioritization of relevant locations in the visual field without moving the eyes—was first noted by Helmholtz (translated in ref. 4), one of the founders of modern perception science. In a dark room, he fixed his gaze at the center of a page of text, attended to "the dark field off to the side," and briefly illuminated the page from behind. He reports that he then "perceived several groups of letters in that [attended] region of the field … The letters in most of the remaining part of the field, however, had not reached perception, not even those that were close to the point of fixation." Since then, psychologists and neuroscientists have puzzled over the dual problem of how attention enhances sensory information and why it apparently cannot enhance all of it.

We now have a substantial body of empirical work showing how attention influences perception and modulates brain activity, with much of this work focused on visual spatial attention. Spatial attention improves even basic visual abilities, which compose the building blocks of vision (1). Attention boosts detection of faint stimuli, increases spatial resolution and acuity, and sharpens our ability to discriminate between similar stimuli. These perceptual improvements are supported by various neural mechanisms (5, 6). Spatial attention increases activity in neural populations specialized for attended locations; reduces the size of spatial receptive fields, enhancing spatial fidelity; reduces correlated noise in neural populations, improving the quality of the visual representation; and shifts population response profiles to better align with the attended visual information. Computational theories of attention have been developed to explain several of these findings. A particularly successful computational framework is the normalization model of attention (7). In this model, attention modulates the gain of neural activity before local contextual modulation, and together these two computational steps determine sensory responses. The normalization model of attention is one example in a history of process models that implement step-by-step operations in which attention changes some aspect of visual processing.

Srivastava et al. (3) took a different approach to model visual spatial attention. Instead of explicitly implementing mechanisms for visual processing and attentional modulation based on known physiology, they trained a convolutional neural network (CNN) to perform a target detection task with a spatial cue and examined the emergent behavior and underlying CNN responses.

The authors started by defining a simple task, the detection of a tilted line that could appear on the left or right of the image. A box, which could also appear on the left or right of the image, served as the spatial cue, with 80% probability that when the target appeared, it was inside the box. This task was similar to those used in human studies in which an informative spatial cue improved the speed and accuracy of target detection (8, 9) and was linked to enhanced visual processing (10–12)—typical effects of spatial attention. Here, the CNN was trained to perform the target detection task by adjusting the network weights in its three convolutional layers and two fully connected layers. In previous work (13), the same authors showed that the trained CNN exhibited better detection accuracy at the cued location compared to the uncued location, similar to human behavior, even though no mechanism for attention had been built into the network.

In the current paper, they asked what neural properties in the trained CNN generated this attention-like behavior. Notably, unlike in neurophysiology studies where only a tiny fraction of the brain's neurons can be studied, Srivastava et al. (3) characterized all 1.8 million model neurons across ten separately trained networks. To do so, they used a clever, neuroscience-inspired methodology to contrast neuronal responses to different probe images, which uncovered neurons tuned to the target, the box cue, or jointly tuned to the target and cue. Such neurons were found most often in the first fully connected "dense" layer and rarely in earlier layers. In the dense layer, some neurons were tuned to the target appearing at one stimulus location (stronger responses to targets than nontargets) and showed a classic cueing effect, with enhanced responses when the box cue surrounded the

Author affiliations: [a]Department of Psychological and Brain Sciences, Boston University, Boston, MA 02215

Author contributions: A.F.C. and R.N.D. wrote the paper.

The authors declare no competing interest.

See companion article, "Emergent neuronal mechanisms mediating covert attention in convolutional neural networks," 10.1073/pnas.2411909122.

[1]To whom correspondence may be addressed. Email: angusc@bu.edu.

Published December 31, 2025.

target. However, other neurons exhibited response properties not previously observed in neurophysiological studies. For example, some "location summation" neurons responded to cues and targets appearing at either location, with no spatial specificity. Meanwhile, "location opponent" neurons were excited by cue–target combinations at one location but inhibited when the cue and target appeared at the other location.

In a new study in PNAS, Srivastava et al. (3) instead investigate how attention-like behavior can emerge from an artificial neural network trained on a visual task.

Importantly, the authors went one step further by asking whether the novel response profiles of CNN model neurons could be found in real data. Comparison with neurophysiological data previously collected from mice superior colliculus (14) revealed previously unreported evidence for both location-summation and location-opponent neurons, supporting the predictions from the CNN. However, whereas the CNN also had neurons that had a mix of summative and opponent properties (e.g., location summation for targets but opponency for cues), such neurons were not observed in the superior colliculus data.

The response profiles recovered for neurons in the trained CNN suggest that attention-like behavior came about through statistical learning of combinations of image features associated with the presence and absence of cues and targets. For example, because targets mostly appeared with cues, neurons tuned to their joint image features emerged during training. Such neurons would be expected to give rise to behavioral cueing effects similar to those seen in the attention literature. In trials with a valid spatial cue, the target appeared inside the box as usual, so the cue-target neurons could be used by the network to solve the detection task, improving performance. But in trials with an invalid spatial cue, the target appeared without the box around it. Then, the cue-target neurons could no longer be used by the network to solve the task in the way it had been trained, so performance was impaired. By "opening the black box" of the CNN, Srivistava et al. provided insight into the mechanisms underlying the attention-like behavior of the model.

Are these mechanisms the same as the ones mediating attention in humans and animals? Attention is a broad domain covering many mechanisms of prioritization (2). The statistical learning–based mechanisms emerging during CNN training might hold explanatory power for predicting neural responses in extensively trained animals—as the authors demonstrated with the superior colliculus data—and in other situations where performance improvement arises from extended learning. In the human literature, "contextual cueing" and "experience-based attention" refer to findings where performance improves after exposure to regularities in the environment (15, 16). Such improvements have usually been attributed to the training of an attentional mechanism that would, for example, shift spatial attention to the location most likely to contain a target, but Srivastava et al.'s (3) results suggest that other explanations tied to changes in neural tuning properties could be further considered.

The mechanisms used by the CNN are unlikely to explain the full range of attentional phenomena, however, for two main reasons. First, statistical learning–based mechanisms lack the flexibility of human attention. When waiting for a rideshare, for example, we can find out the color of the car that is coming to pick us up and immediately use feature-based attention to prioritize that color, without any training. We can even attend to locations in space without knowing what will appear there. And we can prioritize not just based on cues that occur simultaneously with targets, but cues or instructions that occurred earlier. Neurons tuned to cue–target combinations are not in general necessary to allocate attention, and much of our attentional behavior could not be carried out by such neurons alone.

Second, although statistical learning–based mechanisms can lead to performance improvements and impairments in certain situations, they do not have the kinds of processing constraints that lead to true selectivity. For example, humans show performance benefits at attended locations and costs at unattended locations even when location uncertainty is subsequently eliminated by a 100% valid response cue (17–19). However, as noted by the authors, a CNN trained on this task would not be expected to exhibit attention-like behavior. With a fully predictive response cue, there would be no incentive to encode the precue without additional processing bottlenecks. In large part, the cue-driven performance improvements exhibited by the CNN resulted from statistical associations that improved decisions about target presence rather than changes in whether and how sensory information was represented to begin with.

Overall, Srivastava et al. (3) demonstrate that training CNNs on classic cognitive tasks can be a productive avenue to examine the neural mechanisms that support cognitive and decision processes. Excitingly, this study shows how applying neurophysiological analysis strategies to CNNs can enable detailed characterizations of model mechanisms that can then be compared with data from humans and animals performing similar cognitive tasks. Important questions that arise in all efforts to train neural networks on cognitive tasks are whether the model learns the task in the same way humans would and whether the trained model can generalize to a wider range of tasks linked to the cognitive process of interest. Answering such questions in follow-up work will also be necessary for the model Srivastava et al. (3) investigated, but already, the current study provides a thought-provoking new perspective on the kinds of neural mechanisms that can generate behavioral signatures of attention.

1. M. Carrasco, Visual attention: The past 25 years. *Vis. Res.* **51**, 1484–1525 (2011), 10.1016/j.visres.2011.04.012.
2. R. N. Denison, Visual temporal attention from perception to computation. *Nat. Rev. Psychol.* **3**, 264–274 (2024), 10.1038/s44159-024-00294-0.
3. S. Srivastava, W. Y. Wang, M. P. Eckstein, Emergent neuronal mechanisms mediating covert attention in convolutional neural networks. *Proc. Natl. Acad. Sci. U.S.A.* 122, e2411909122 (2025).
4. K. Nakayama, M. Mackeben, Sustained and transient components of focal visual attention. *Vis. Res.* **29**, 1631–1647 (1989), 10.1016/0042-6989(89)90144-2.

5. J. H. R. Maunsell, Neuronal mechanisms of visual attention. *Annu. Rev. Vis. Sci.* **1**, 373–391 (2015), 10.1146/annurev-vision-082114-035431.
6. D. A. Ruff, A. M. Ni, M. R. Cohen, Cognition as a window into neuronal population space. *Annu. Rev. Neurosci.* **41**, 77–97 (2018), 10.1146/annurev-neuro-080317-061936.
7. J. H. Reynolds, D. J. Heeger, The normalization model of attention. *Neuron* **61**, 168–185 (2009), 10.1016/j.neuron.2009.01.002.
8. M. I. Posner, Orienting of attention. *Q. J. Exp. Psychol.* **32**, 3–25 (1980), 10.1080/00335558008248231.
9. S. Ling, M. Carrasco, Sustained and transient covert attention enhance the signal via different contrast response functions. *Vis. Res.* **46**, 1210–1220 (2006), 10.1016/j.visres.2005.05.008.
10. S. A. Hillyard, L. Anllo-Vento, Event-related brain potentials in the study of visual selective attention. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 781–787 (1998), 10.1073/pnas.95.3.781.
11. S. Kastner, M. A. Pinsk, P. De Weerd, R. Desimone, L. G. Ungerleider, Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron* **22**, 751–761 (1999), 10.1016/S0896-6273(00)80734-5.
12. M. A. Silver, D. Ress, D. J. Heeger, Neural correlates of sustained spatial attention in human early visual cortex. *J. Neurophysiol.* **97**, 229–237 (2007), 10.1152/jn.00677.2006.
13. S. Srivastava, W. Y. Wang, M. P. Eckstein, Emergent human-like covert attention in feedforward convolutional neural networks. *Curr. Biol.* **34**, 579–593 (2024), 10.1016/j.cub.2023.12.058.
14. L. Wang, J. P. Herman, R. J. Krauzlis, Neuronal modulation in the mouse superior colliculus during covert visual selective attention. *Sci. Rep.* **12**, 2482 (2022), 10.1038/s41598-022-06410-5.
15. M. M. Chun, Y. Jiang, Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cogn. Psychol.* **36**, 28–71 (1998), 10.1006/cogp.1998.0681.
16. J. J. Geng, M. Behrmann, Spatial probability as an attentional cue in visual search. *Percept. Psychophys.* **67**, 1252–1268 (2005), 10.3758/BF03193557.
17. B. A. Dosher, Z.-L. Lu, Noise exclusion in spatial attention. *Psychol. Sci.* **11**, 139–146 (2000), 10.1111/1467-9280.00229.
18. A. Fernández, H.-H. Li, M. Carrasco, How exogenous spatial attention affects visual representation. *J. Vis.* **19**, 4 (2019), 10.1167/19.11.4.
19. S. Ling, M. Carrasco, When sustained attention impairs perception. *Nat. Neurosci.* **9**, 1243–1245 (2006), 10.1038/nn1761.