NEUROSCIENCE

Using neuroscience to develop artificial intelligence

Combining deep learning with brain-like innate structures may guide network models toward human-like learning

By **Shimon Ullman**

When the mathematician Alan Turing posed the question "Can machines think?" in the first line of his seminal 1950 paper that ushered in the quest for artificial intelligence (AI) (*1*), the only known systems carrying out complex computations were biological nervous systems. It is not surprising, therefore, that scientists in the nascent field of AI turned to brain circuits as a source for guidance. One path that was taken since the early attempts to perform intelligent computation by brain-like circuits (*2*), and which led recently to remarkable successes, can be described as a highly reductionist approach to model cortical circuitry. In its basic current form, known as a "deep network" (or deep net) architecture, this brain-inspired model is built from successive layers of neuron-like elements, connected by adjustable weights, called "synapses," after their biological counterparts (*3*). The application of deep nets and related methods to AI systems has been transformative. They proved superior to previously known methods in central areas of AI research, including computer vision, speech recognition and production, and playing complex games. Practical applications are already in broad use, in areas such as computer vision and speech and text translation, and large-scale efforts are under way in many other areas. Here, I discuss how additional aspects of brain circuitry could supply cues for guiding network models toward broader aspects of cognition and general AI.

The key problem in deep nets is learning, which is the adjustment of the synapses to produce the desired outputs to their input patterns. The adjustment is performed automatically based on a set of training examples, which are provided by input patterns coupled with their desired outputs. The learning process then adjusts the weights to produce the desired outputs to the training input patterns. Successful learning will cause the network to go beyond memorizing the training examples, and be able to generalize, and provide correct outputs to new input patterns, which were not encountered during the learning process.

Comparisons of deep network models with empirical physiological, functional magnetic resonance imaging, and behavioral data have shown some intriguing similarities between brains and the new models (*4*), as well as dissimilarities (*5*) (see the figure). In comparisons with the primate visual system, similarities between physiological and model responses were closer for the early compared with later parts of the neuronal responses, suggesting that the deep network models may capture better the early processing stages, compared with later, more cognitive stages.

In addition to deep nets, AI models recently incorporated another major aspect of brain-like computations: the use of reinforcement learning (RL), where reward signals in the brain are used to modify behavior. Brain mechanisms involved in this form of learning have been studied extensively (*6*), and computational models (*7*) have been used in areas of AI, in particular in robotics applications. RL is used in the context of an agent (a person, animal, or robot) behaving in the world, and receiving reward signals in return. The goal is to learn an optimal "policy," which is a mapping from states to actions, so as to maximize an overall measure of the reward obtained over time. RL methods have been combined in recent AI algorithms with deep network methods, applied in particular to game playing, ranging from popular video games to highly complex games such as chess, Go, and shogi. Combining deep nets with RL produced stunning results in game playing, including convincing defeats of the world's top Go players, or reaching a world-champion level in chess after ~4 hours of training, starting from just the rules of the game, and learning from games played internally against itself (*8*).

From the standpoint of using neuroscience to guide AI, this success is surprising, given the highly reduced form of the network models compared with cortical circuitry. Some additional brain-inspired aspects, for example, normalization across neuronal groups, or the use of spatial attention, have been incorporated into deep network models, but in general, almost everything that we know about neurons—their structure, types, interconnectivity, and so on—was left out of deep-net models in their current form. It is currently unclear which aspects of the biological circuitry are computationally essential and could also be useful for network-based AI systems, but the differences in structure are prominent. For example, biological neurons are highly complex and diverse in terms of their morphology, physiology, and neurochemistry. The inputs to a typical excitatory pyramidal neuron are distributed over complex, highly branching basal and apical dendritic trees. Inhibitory cortical neurons come in a variety of different morphologies, which are likely to perform different functions. None of this heterogeneity and other complexities are included in typical deep-net models, which use instead a limited set of highly simplified homogeneous artificial neurons. In terms of connectivity between units in the network, cortical circuits in the brain are more complex than current deep network models and include rich lateral connectivity between neurons in the same layer, by both local and long-range connections, as well as top-down connections going from high to low levels of the hierarchy of cortical regions, and possibly organized in typical local "canonical circuits."

The notable successes of deep network–based learning methods, primarily in problems related to real-world perceptual data such as vision and speech, have recently been followed by increasing efforts to confront problems that are more cognitive in nature. For example, in the domain of vision, network models were developed initially to deal with perceptual problems such as object classification and segmentation. Similar methods, with some extensions, are now being applied to higher-level problems such as image captioning, where the task is to produce a short verbal description of an image, or to the domain of visual question answering, where the task is to produce adequate answers to queries posed in natural language (that is, human communication) about the content of an image. Other, nonvisual tasks include judging humor, detecting sarcasm, or capturing aspects of intuitive physics or social understanding. Similar methods are also being developed for challenging real-world applications such as online translation, flexible personal assistants, medical diagnosis, advanced robotics, or automatic driving.

With these large research efforts, and the

Department of Computer Science, Weizmann, Institute of Science, Rehovot, Israel. Email: shimon.ullman@weizmann.ac.il

huge funds invested in future AI applications, a major open question is the degree to which current approaches will be able to produce "real" and human-like understanding, or whether additional, perhaps radically different, directions will be needed to deal with broad aspects of cognition, and artificial general intelligence (AGI) (*9*, *10*). The answers to this question are unknown, and the stakes are high, both scientifically and commercially.

If the success of current deep network models in producing human-like cognitive abilities proves to be limited, a natural place to look for guidance is again neuroscience. Can aspects of brain circuitry, overlooked in AI models so far, provide a key to AGI? Which aspects of the brain are likely to be particularly important? There are at present no obvious answers, because our understanding of cortical circuitry is still limited, but I will briefly discuss a general aspect by which brains and deep network models appear to be fundamentally different and that is likely to have an important functional role in the quest for human-like AGI. The difference centers on the age-old question about the balance between empiricism and nativism in cognition, namely, the relative roles of innate cognitive structures and general learning mechanisms. Current AI modeling leans heavily toward the empiricist side, using relatively simple and uniform network structures, and relying primarily on extended learning, using large sets of training data. By contrast, biological systems often accomplish complex behavioral tasks with limited training, building upon specific preexisting network structures already encoded in the circuitry prior to learning. For example, different animal species, including insects, fish, and birds, can perform complex navigation tasks relying in part on an elaborate set of innate domain-specific mechanisms with sophisticated computational capabilities. In humans, infants start to develop complex perceptual and cognitive skills in the first months of life, with little or no explicit training. For example, they spontaneously recognize complex objects such as human hands, follow other peoples' direction of gaze, and distinguish visually whether animated characters are helping or hindering others, and a variety of other tasks, which exhibit an incipient understanding of physical and social interactions. A large body of developmental studies have suggested that this fast, unsupervised learning is possible because the human cognitive system is equipped, through evolution, with basic innate structures that facilitate the acquisition of meaningful concepts and cognitive skills (*11*, *12*).

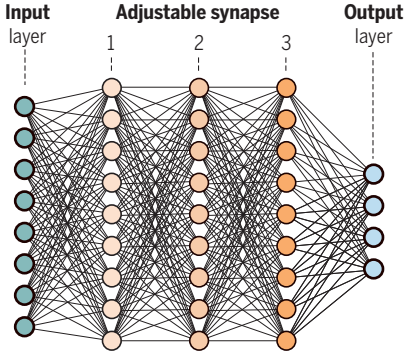
The superiority of human cognitive learning and understanding compared with existing deep network models may largely result from the much richer and complex innate structures incorporated in the human cognitive system. Recent modeling of visual learning in infancy (*13*) has shown a useful combination of learning and innate mechanisms, where meaningful complex concepts are neither innate nor learned on their own. The innate components in this intermediate view are not developed concepts, but simpler "proto concepts," which provide internal teaching signals and guide the learning system along a path that leads to the progressive acquisition and organization of complex concepts, with little or no explicit training. For example, it was shown how a particular pattern of image motion can provide a reliable internal teaching signal for hand recognition. The detection of hands, and their engagement in object manipulation, can in turn guide the learning system toward detecting direction of gaze, and detecting gaze targets is known to play a role in learning to infer people's goals (*14*). Such innate structures could be implemented by an arrangement of local cortical regions with specified initial connectivity, supplying inputs and error signals to specific targets.

Useful preexisting structures could also be adopted in artificial network models to make their learning and understanding more human-like. The challenge of discovering useful preexisting structures can be approached by either understanding and mimicking related brain mechanisms, or by developing computational learning methods that start "from scratch" and discover structures that support an agent, human or artificial, that learns to understand its environment in an efficient and flexible manner. Some attempts have been made in this direction (*15*), but in general, the computational problem of "learning innate structures" is different from current learning procedures, and it is poorly understood. Combining the empirical and computational approaches to the problem is likely to benefit in the long run both neuroscience and AGI, and could eventually be a component of a theory of intelligent processing that will be applicable to both. ■

Brain circuitry and learning

A major open question is whether the highly simplified structures of current network models compared with cortical circuits are sufficient to capture the full range of human-like learning and cognition.

Complex neural network
Connectivity in cortical networks includes rich sets of connections, including local and long-range lateral connectivity, and top-down connections from high to low levels of the hierarchy.



Informed AI network
Biological innate connectivity patterns provide mechanisms that guide human cognitive learning. Discovering similar mechanisms, by machine learning or by mimicking the human brain, may prove crucial for future artificial systems with human-like cognitive abilities.

GRAPHIC: C. BICKEL/*SCIENCE*

REFERENCES AND NOTES

1. A. M. Turing, *Mind* **59**, 433 (1950).
2. F. Rosenblatt, *Psychol. Rev.* **65**, 386 (1958).
3. Y. LeCun *et al.*, *Nature* **521**, 436 (2015).
4. N. J. Majaj *et al.*, *J. Neurosci.* **35**, 13402 (2015).
5. R. Rajalingham *et al.*, *J. Neurosci.* **38**, 7255 (2018).
6. D. Lee *et al.*, *Annu. Rev. Neurosci.* **35**, 287 (2012).
7. R. S. Sutton *et al.*, *Reinforcement Learning: An Introduction* (MIT Press, 1998).
8. D. Silver *et al.*, *Nature* **550**, 354 (2017).
9. D. Hassabis *et al.*, *Neuron* **95**, 245 (2017).
10. B. M. Lake *et al.*, *Behav. Brain Sci.* **40**, e253 (2017).
11. E. S. Spelke, K. D. Kinzler, *Dev. Sci.* **10**, 89 (2007).
12. S. Carey, *The Origin of Concepts* (Oxford Univ. Press, New York, 2009).
13. S. Ullman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 18215 (2012).
14. A. T. Phillips *et al.*, *Cognition* **85**, 53 (2002).
15. E. Real *et al.*, *Proc. 34th Int. Conf. Machine Learning, PMLR* **70**, 2902 (2017).

ACKNOWLEDGMENTS

Supported by European Union's Horizon 2020 Framework 785907 (HBP SGA2). S. U. thanks colleagues at the Center for Minds, Brains and Machines at Massachusetts Institute of Technology for helpful discussions.

10.1126/science.aau6595