VIEWPOINT

Deep Learning—A Technology With the Potential to Transform Health Care

Geoffrey Hinton, PhD
Google Brain Team and Department of Computer Science, University of Toronto, Ontario, Canada.

Viewpoint and Editorial

Widespread application of artificial intelligence in health care has been anticipated for half a century. For most of that time, the dominant approach to artificial intelligence was inspired by logic: researchers assumed that the essence of intelligence was manipulating symbolic expressions, using rules of inference. This approach produced expert systems and graphical models that attempted to automate the reasoning processes of experts. In the last decade, however, a radically different approach to artificial intelligence, called deep learning, has produced major breakthroughs and is now used on billions of digital devices for complex tasks such as speech recognition, image interpretation, and language translation. The purpose of this Viewpoint is to give health care professionals an intuitive understanding of the technology underlying deep learning. In an accompanying Viewpoint, Naylor[1] outlines some of the factors propelling adoption of this technology in medicine and health care.

What Neural Networks Can Do

Artificial neural networks are inspired by the ability of brains to learn complicated patterns in data by changing the strengths of synaptic connections between neurons.

> Artificial neural networks are inspired by the ability of brains to learn complicated patterns in data by changing the strengths of synaptic connections between neurons.

Deep learning uses deep networks with many intermediate layers of artificial "neurons" between the input and the output, and, like the visual cortex, these artificial neurons learn a hierarchy of progressively more complex feature detectors. By learning feature detectors that are optimized for classification, deep learning can substantially outperform systems that rely on features supplied by domain experts or that are designed by hand.[2]

Deep learning excels at modeling extremely complicated relationships between inputs and outputs. This technology can be used for tasks as different as predicting future medical events from past events[3] and predicting cardiovascular health from fundus images of the retina.[4] Deep learning is already achieving results that equal or surpass those of human experts. For example, in a 2017 report, Esteva et al[5] compiled a database of 129 450 labeled images of hundreds of different skin lesions. Approximately 2000 images with accurate diagnostic labels based on skin biopsies were used for test purposes, and the rest were used to retrain a convolutional neural network that had previously been trained to recognize everyday objects in cluttered images. The skin lesion images used for retraining varied widely in quality, and no further information was provided to the convolutional neural network other than the image pixels and the lesion label. The network and groups of 21 to 25 board-certified dermatologists then reviewed subsets of the unlabeled test images and decided whether the correct clinical course was a biopsy for possible malignancy or reassurance of the patient. Sensitivity for the majority of the dermatologists was lower than that of the convolutional neural network when matched for specificity, and their specificity was lower than that of the convolutional neural network when matched for sensitivity for identifying images with melanoma, as well as for images of basal and squamous cell carcinoma.

A Brief History of Artificial Neural Networks

The simple neural nets of the 1960s had to be provided with hand-designed feature detectors and they simply learned how much weight to give to each detector. The introduction in 1986 of the back-propagation procedure[6] (explained below) allowed neural networks to design their own feature detectors, which made them much more powerful at modeling complicated relationships between their inputs and outputs, especially when they used multiple layers of learned features. However, despite some promising results in the 1990s in reading the numeric amounts on checks, it proved difficult to train deep neural networks and they did not consistently outperform other simpler machine-learning techniques.

What changed? In simple terms, computers became millions of times faster, data sets got thousands of times bigger, and researchers discovered many technical refinements that made neural nets easier to train.

How Deep Learning Works

Consider the problem of deciding whether a patient has a specific disease when given a large number of numeric input variables that represent characteristics of the patient. One standard approach is to use simple logistic regression that estimates how to weight each input variable so that their weighted sum is a good indicator of the disease. Since health and disease often involve complex interactions, a statistician can add extra inputs, known as interaction terms, each representing the product of 2 or more input variables. However, if multiway interactions need to be modeled, the number of interaction terms increases exponentially.

Corresponding Author: Geoffrey Hinton, PhD, Google Brain Team and Department of Computer Science, University of Toronto, 6 King's College Rd, Toronto, ON M5S 1A1, Canada (geoffrey.hinton@gmail.com).

The neural network alternative is to add a layer of "hidden factors" (ie, features). The first step is to determine which hidden factors are active, and then the active ones are used to determine whether the disease is present. To prevent the model from becoming too big while allowing factors to reflect many input variables, the *number* of hidden factors is limited, rather than the number of input variables that contribute to each factor. The challenge is then to learn a good set of hidden factors by repeatedly modifying the weights on connections from the input variables to the hidden factors and the weights on connections from the hidden factors to the output variable.

In principle, a learning procedure could repeatedly choose single weights at random, make a small change, and keep this change if it improves the performance of the whole net, but this would be extremely slow. In a neural net with a million weights, back-propagation achieves the same goal about a million times faster than blind trial and error. Instead of changing weights and measuring the effect, the neural network takes the discrepancy between the output produced by the network for each patient and the target output and propagates this discrepancy backward through the network to compute, for all of the weights, how a small change in a weight would reduce the discrepancy. The network then changes every weight in the direction that reduces the discrepancy by an amount proportional to how rapidly it reduces the discrepancy.

Back-propagation can be used to train deep networks that have many layers of hidden factors, with each layer of features depending on the features in the preceding layer. For complex image interpretation, as occurs in many medical applications, neural networks can be improved by making a separate copy of each feature detector for every position in the image. After updating of the incoming weights of each copy, the corresponding weights are averaged so that all copies use an identical set of weights. This is called a convolutional neural network,[7] and it allows knowledge acquired by looking at one part of an image to be applied at every location in subsequent images.

For modeling *sequences*, such as a patient's medical history, a "recurrent" neural network can be used that takes in one term at a time.[7] In addition to the connections coming from the layer below, each layer of a recurrent network has weighted connections coming from its own activations at the previous time step, and this allows the layers to accumulate and transform information over time. The back-propagation phase then sends the discrepancy between a prediction of the next term in the sequence and the actual next term backward through the layers and also backward through the time steps.

A Caveat About Interpretability

Understandably, clinicians, scientists, patients, and regulators would all prefer to have a simple explanation for how a neural net arrives at its classification of a particular case. In the example of predicting whether a patient has a disease, they would like to know what hidden factors the network is using. However, when a deep neural network is trained to make predictions on a big data set, it typically uses its layers of learned, nonlinear features to model a huge number of complicated but weak regularities in the data. It is generally infeasible to interpret these features because their meaning depends on complex interactions with uninterpreted features in other layers. Also, if the same neural net is refit to the same data, but with changes in the initial random values of the weights, there will be different features in the intermediate layers. This reflects that unlike models in which an expert specifies the hidden factors, a neural net has many different and equally good ways of modeling the same data set. It is not trying to identify the "correct" hidden factors. It is merely using hidden factors to model the complicated relationship between the input variables and the output variables.

The Future of Deep Learning

As data sets get bigger and computers become more powerful, the results achieved by deep learning will get better, even with no improvement in the basic learning techniques, although these techniques are being improved. The neural networks in the human brain learn from fewer data and develop a deeper, more abstract understanding of the world. In contrast to machine-learning algorithms that rely on provision of large amounts of labeled data, human cognition can find structure in unlabeled data, a process commonly termed *unsupervised learning*. The creation of a smorgasbord of complex feature detectors based on unlabeled data appears to set the stage for humans to learn a classifier from only a small amount of labeled data. How the brain does this is still a mystery, but will not remain so. As new unsupervised learning algorithms are discovered, the data efficiency of deep learning will be greatly augmented in the years ahead, and its potential applications in health care and other fields will increase rapidly.

ARTICLE INFORMATION

Published Online: August 30, 2018. doi:10.1001/jama.2018.11100

Conflict of Interest Disclosures: The author has completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr Hinton reports owning stock in Google.

Additional Contributions: I am deeply indebted to C. David Naylor, MD, DPhil, for discussions that shaped this Viewpoint and for sharing his knowledge of health care.

REFERENCES

1. Naylor CD. On the prospects for a (deep) learning health care system [published online August 30, 2018]. *JAMA*. doi:10.1001/jama.2018.11103

2. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

3. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*. 2018;1(1):18. doi:10.1038/s41746-018-0029-1

4. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2:158-164. doi:10.1038/s41551-018-0195-0

5. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056

6. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986;323(6088):533. doi:10.1038/323533a0

7. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Vol 1. Cambridge, MA: MIT Press; 2016.