Depths of learning

In the past five years, neural networks associated with the name 'deep learning' have taken centre stage in machine learning research. They handle many common tasks as well as or better than people, including speech and image recognition, and language translation. They're increasingly being used to do anything from interpreting X-ray images to managing investments, and even moving on to influence scientific research. In a recent study (A. B. Farimani, J. Gomes and V. S. Pande, preprint at https://arxiv.org/abs/1709.02432; 2017), researchers used such algorithms to build mathematical theories of complex processes of heat conduction or fluid flow directly from experimental data, with no human theorist involved.

Are we only a few years and minor conceptual advances away from true human-level artificial intelligence? Or is the gap still large? Commercial interests hype the potential for artificial intelligence, and it is likely to have many profound economic consequences. Yet experts in machine learning are more cautious, noting that these neural networks and the algorithms they embody aren't nearly as powerful as they sometimes appear.

For example, subtle changes to an image, so insignificant that no human would even notice, can make an algorithm see something that isn't there — an ordinary tabby cat becomes a bowl of guacamole, or a parking sign a refrigerator. It's an open question whether these weaknesses can be patched, as researchers have not succeeded even in ensuring that algorithms can detect a spoofed example when it encounters one. In a recent study, computer scientists Nicholas Carlini and David Wagner of the University of California, Berkeley, tested ten detection schemes proposed over the past year and found that they could all be evaded (preprint at https://arxiv.org/abs/1705.07263; 2017).

So deep learning algorithms in their current form appear to be quite fragile, as does our understanding of why they work when they do. As they learn on training data, some algorithms go through two distinct stages — one akin to memorization of the data, and a second involving compression, as the algorithm reduces dimension by letting go of details in the data that are irrelevant to the classification problem. But other algorithms don't show the two stages, and actually perform better if researchers step in and limit their training time. Why remains unknown.

Given such mysteries, some researchers think our understanding of deep learning algorithms may still be quite primitive, and that the rapid progress seen in recent years may not continue, as the field hits a wall. In a recent article, computer scientist Gary Marcus of New York University explores ten key problems facing deep learning methods right now that should caution against excessive expectations for artificial intelligence (preprint at https://arxiv.org/abs/1801.00631; 2018).

Deep learning algorithms in their current form appear to be quite fragile, as does our understanding of why they work when they do.

One of his points is about algorithmic inflexibility. Deep learning neural networks generally train with huge amounts of data to tackle well-defined and unchanging problems. Translate some text. Recognize an object in a photo. What these algorithms need for a given problem is provided in a neat package. While this includes some problems of human relevance, it excludes a huge range of other tasks. What's a good way to get a tangled rope out of a bicycle wheel? People handle tasks like this all the time, without being given any hints about the nature of an answer, or what information might be needed to approach it.

People also learn abstract relationships in only a few trials. Suppose I define 'schmister' — following another example given by Marcus — as a sister aged between 10 and 21 years. Any person would quickly be able to tell if he or she had any schmisters. Deep learning can't do this at all. Apple's Siri and similar algorithms are hopeless at learning even the simplest abstractions through verbal definition. Most of the things we do on a daily basis may be unsuitable for algorithms, at least of this kind.

Related to this inflexibility is a lack of conceptual depth. The term deep learning refers to the many layers of connections between inputs and outputs used in a these neural networks, in contrast to earlier networks that used only a few layers. Deep learning isn't really deep in thinking terms, as these networks utterly lack any ability to deal in abstract concepts. Unlike any human infant, they can't learn to understand concepts like 'shiny' or 'fairness'.

Another problem, Marcus notes, has to do with what logicians call 'open-ended inference'. Humans easily see the difference between 'John promised Mary to leave' and 'John promised to leave Mary', and can draw conclusions from a phrase by using information that isn't explicitly included in that phrase. We draw on background knowledge about the entities involved. Deep learning algorithms have no such background knowledge, and generally can give answers that only involve explicitly given information.

A more alarming problem, especially for applications, is reliability, and the ease with which algorithms can be spoofed. A stop sign dusted with a small bit of noise will look to an algorithm like a speed limit sign. A 3D-printed toy turtle, its colours tweaked appropriately, will be judged to be a rifle. These errors are amusing in research, but potentially deadly in applications — if algorithms were entrusted to run air traffic control systems, for example, or to scan luggage for explosives.

Marcus also mentions one final problem that may hamper broad applications of such algorithms: transparency, or a lack if it. We build aircraft engines out of many small known parts, and guarantee performance by understanding how those parts interact. Today's algorithms come with none of this transparency. Deep learning algorithms are still block boxes, created through a process of learning and evolution, their structures described by billions of parameters and often unknown even to their creators. Any such algorithm applied in areas like financial trading or medical diagnosis ought to be open to human understanding, so human users can inquire about why a system made a particular decision.

We may well be entering an era of artificial intelligence. But, in the near future, that intelligence is likely to be of a rather inhuman kind, and have less impact than many expect. ❐

Mark Buchanan

Published online: 6 April 2018
https://doi.org/10.1038/s41567-018-0098-8

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.