TECHNOLOGY FEATURE

DEEP LEARNING FOR BIOLOGY

A popular artificial-intelligence method provides a powerful tool for surveying and classifying biological data. But for the uninitiated, the technology poses significant difficulties.

ALFRED PASIEKA/SPL/GETTY

The brain's neural network has long inspired artificial-intelligence researchers.

BY SARAH WEBB

Four years ago, scientists from Google showed up on neuroscientist Steve Finkbeiner's doorstep. The researchers were based at Google Accelerated Science, a research division in Mountain View, California, that aims to use Google technologies to speed scientific discovery. They were interested in applying 'deep-learning' approaches to the mountains of imaging data generated by Finkbeiner's team at the Gladstone Institute of Neurological Disease in San Francisco, also in California.

Deep-learning algorithms take raw features from an extremely large, annotated data set, such as a collection of images or genomes, and use them to create a predictive tool based on patterns buried inside. Once trained, the algorithms can apply that training to analyse other data, sometimes from wildly different sources.

The technique can be used to "tackle really hard, tough, complicated problems, and be able to see structure in data — amounts of data that are just too big and too complex for the human brain to comprehend", Finkbeiner says.

He and his team produce reams of data using a high-throughput imaging strategy known as robotic microscopy, which they had developed for studying brain cells. But the team couldn't analyse its data at the speed it acquired them, so Finkbeiner welcomed the opportunity to collaborate.

"I can't honestly say at the time that I had a clear grasp of what questions might be addressed with deep learning, but I knew that we were generating data at about twice to three times the rate we could analyse it," he says.

Today, those efforts are beginning to pay off. Finkbeiner's team, with scientists at Google, trained a deep algorithm with two sets of cells, one artificially labelled to highlight features that scientists can't normally see, the other unlabelled. When they later exposed the algorithm to images of unlabelled cells that it had never seen before, Finkbeiner says, "it was astonishingly good at predicting what the labels should be for those images". A publication detailing that work is now in the press.

Finkbeiner's success highlights how deep learning, one of the most promising branches of artificial intelligence (AI), is making inroads in biology. The algorithms are already infiltrating modern life in smartphones, smart speakers and self-driving cars. In biology, deep-learning algorithms dive into data in ways that humans can't, detecting features that might otherwise be impossible to catch. Researchers are using the algorithms to classify cellular images, make genomic connections, advance drug discovery and even find links across different data types, from genomics and imaging to electronic medical records.

More than 440 articles on the bioRxiv preprint server discuss deep learning; PubMed lists more than 700 references in 2017. And the tools are on the cusp of becoming widely available to biologists and clinical researchers. But researchers face challenges in understanding just what these algorithms are doing, and ensuring that they don't lead users astray.

TRAINING SMART ALGORITHMS

Deep-learning algorithms (see 'Deep thoughts') rely on neural networks, a computational model first proposed in the 1940s, ▶

in which layers of neuron-like nodes mimic how human brains analyse information. Until about five years ago, machine-learning algorithms based on neural networks relied on researchers to process the raw information into a more meaningful form before feeding it into the computational models, says Casey Greene, a computational biologist at the University of Pennsylvania in Philadelphia. But the explosion in the size of data sets — from sources such as smartphone snapshots or large-scale genomic sequencing — and algorithmic innovations have now made it possible for humans to take a step back. This advance in machine learning — the 'deep' part — forces the computers, not their human programmers, to find the meaningful relationships embedded in pixels and bases. And as the layers in the neural network filter and sort information, they also communicate with each other, allowing each layer to refine the output from the previous one.

Eventually, this process allows a trained algorithm to analyse a new image and correctly identify it as, for example, Charles Darwin or a diseased cell. But as researchers distance themselves from the algorithms, they can no longer control the classification process or even explain precisely what the software is doing. Although these deep-learning networks can be stunningly accurate at making predictions, Finkbeiner says, "it's still challenging sometimes to figure out what it is the network sees that enables it to make such a good prediction".

Still, many subdisciplines of biology, including imaging, are reaping the rewards of those predictions. A decade ago, software for automated biological-image analysis focused on measuring single parameters in a set of images. For example, in 2005, Anne Carpenter, a computational biologist at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, released an open-source software package called CellProfiler to help biologists to quantitatively measure individual features: the number of fluorescent cells in a microscopy field, for example, or the length of a zebrafish.

But deep learning is allowing her team to go further. "We've been shifting towards measuring things that biologists don't realize they want to measure out of images," she says. Recording and combining visual features such as DNA staining, organelle texture and the quality of empty spaces in a cell can produce thousands of 'features', any one of which can reveal fresh insights. The current version of CellProfiler includes some deep-learning elements, and her team expects to add more-sophisticated deep-learning tools in the next year.

"Most people have a hard time wrapping their heads around this," Carpenter says, "but there's just as much information, in fact maybe more, in a single image of cells as there is in a transcriptomic analysis of a cell population."

That type of processing allows Carpenter's team to take a less supervised approach to translating cell images into disease-associated phenotypes — and to capitalize on it. Carpenter is a scientific adviser to Recursion Pharmaceuticals in Salt Lake City, Utah, which is using its deep-learning tools to target rare, single-gene disorders for drug development.

MINING GENOMIC DATA

When it comes to deep learning, not just any data will do. The method often requires massive, well-annotated data sets. Imaging data provide a natural fit, but so, too, do genomic data.

One biotech firm that is using such data is Verily Life Sciences (formerly Google Life Sciences) in San Francisco. Researchers at Verily — a subsidiary of Google's parent company, Alphabet — and Google have developed a deep-learning tool that identifies a common type of genetic variation, called single-nucleotide polymorphisms, more accurately than conventional tools. Called DeepVariant, the software translates genomic information into image-like representations, which are then analysed as images (see 'Tools for deep diving'). Mark DePristo, who heads deep-learning-based genomic research at Google, expects DeepVariant to be particularly useful for researchers studying organisms outside the mainstream — those with low-quality reference genomes and high error rates in identifying genetic variants. Working with DeepVariant in plants, his colleague Ryan Poplin has achieved error rates closer to 2% than the more-typical 20% of other approaches.

Brendan Frey, chief executive of the Canadian company Deep Genomics in Toronto, also focuses on genomic data, but with the goal of predicting and treating disease. Frey's academic team at the University of Toronto developed algorithms trained on genomic and transcriptomic data from healthy cells. Those algorithms built predictive models of RNA-processing events such as splicing, transcription and polyadenylation within those data. When applied to clinical data, the algorithms were able to identify mutations and flag them as pathogenic, Frey says, even though they'd never seen clinical data. At Deep Genomics, Frey's team is using the same tools to identify and target the disease mechanisms that the software uncovered, to develop therapies derived from short nucleic-acid sequences.

Another discipline with massive data sets that are amenable to deep learning is drug discovery. Here, deep-learning algorithms are helping to solve categorization challenges, sifting through such molecular features as shape and hydrogen bonding to identify criteria on which to rank those potential drugs. For instance, Atomwise, a biotech company based in San Francisco, has developed algorithms that convert molecules into grids of 3D pixels, called voxels. This representation allows the company to account for the 3D structure of proteins and small molecules with atomic precision, modelling features such as the geometries of carbon atoms. Those features are then translated into mathematical vectors that the algorithm can use to predict which small molecules are likely to interact with a given protein, says Abraham Heifets, the company's chief executive. "A lot of the work we do is for [protein] targets with no known binders," he says.

Atomwise is using this strategy to power its new AI-driven molecular-screening programme, which scans a library of 10 million compounds to provide academic researchers with up to 72 potential small-molecule binders for their protein of interest.

Tools for deep diving

Deep-learning tools are evolving rapidly, and labs will need dedicated computational expertise, collaborations or both to take advantage of them.

First, take a colleague with deep-learning expertise out to lunch and ask whether the strategy might be useful, advises Steve Finkbeiner, a neuroscientist at the Gladstone Institutes in San Francisco, California. With some data sets, such as imaging data, an off-the-shelf program might work; for more complicated projects, consider a collaborator, he says. Workshops and meetings can provide training opportunities.

Access to cloud-computing resources means that researchers might not need an on-site computer cluster to use deep learning — they can run the computation elsewhere. Google's TensorFlow, an open-source platform for building deep-learning algorithms, is available on the software-sharing site GitHub, as is an open-source version of DeepVariant, a tool for accurately identifying genetic variation.

Google Accelerated Science, a Google research division based in Mountain View, California, collaborates with a range of scientists, including biologists, says Michelle Dimon, one of its research scientists. Projects require a compelling biological question, large amounts of high-quality, labelled data, and a challenge that will allow the company's machine-learning experts to make unique computational contributions to the field, Dimon says.

Those wishing to get up to speed on deep learning should check out the 'deep review', a comprehensive, crowdsourced review led by computational biologist Casey Greene of the University of Pennsylvania in Philadelphia (T. Ching *et al.* Preprint at bioRxiv http://doi.org/gbpvh5; 2018). S.W.

DEEP THOUGHTS

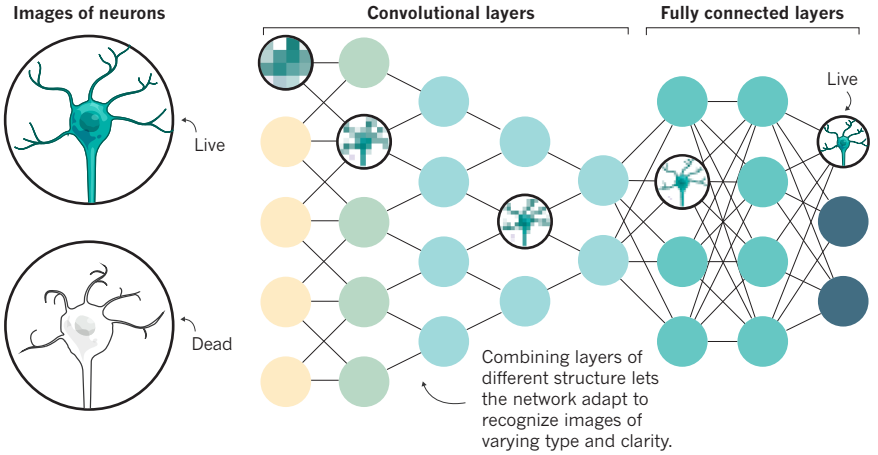
Deep-learning algorithms take many forms. Steve Finkbeiner's lab used a convolutional neural network (CNN) such as this one to identify, with high accuracy, dead neurons in a population of live and dead cells.

INPUT

The network is trained using several hundred thousand annotated images of live and dead cells.

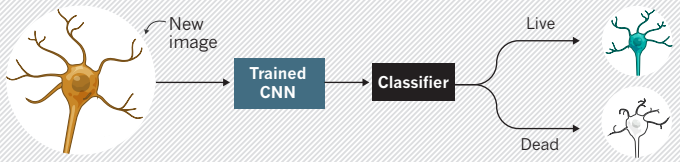
TRAINING AI

Over multiple iterations, the network discovers patterns in the data that can distinguish live from dead cells. Convolutional layers identify structural features of the images, which are integrated in fully connected layers.



APPLICATION

Challenged with unlabelled images, the network assigns each cell as alive or dead with high accuracy.



SOURCE: JEREMY LINSLEY/DREW LINSLEY/STEVE FINKBEINER/THOMAS SERRE

Deep-learning tools could also help researchers to stratify disease types, understand disease subpopulations, find new treatments and match them with the appropriate patients for clinical testing and treatment. Finkbeiner, for instance, is part of a consortium called Answer ALS, an effort to combine a range of data — genomics, transcriptomics, epigenomics, proteomics, imaging and even pluripotent stem-cell biology — from 1,000 people with the neurodegenerative disease amyotrophic lateral sclerosis (also called motor neuron disease). "For the first time, we'll have a data set where we can apply deep learning and look at whether deep learning can uncover a relationship between the things we can measure in a dish around a cell, and what's happening to that patient," he says.

CHALLENGES AND CAUTIONS

For all its promise, deep learning poses significant challenges, researchers warn. As with any computational-biology technique, the results that arise from algorithms are only as good as the data that go in. Overfitting a model to its training data is also a concern. In addition, for deep learning, the criteria for data quantity and quality are often more rigorous than some experimental biologists might expect.

Deep-learning algorithms have required extremely large data sets that are well annotated so that the algorithms can learn to distinguish features and categorize patterns. Larger, clearly labelled data sets — with millions of data points representing different experimental and physiological conditions — give researchers the most flexibility for training an algorithm. Finkbeiner notes that algorithm training in his work improves significantly after about 15,000 examples. Those high-quality 'ground truth' data can be exceptionally hard to come by, says Carpenter.

To circumvent this challenge, researchers have been working on ways to train more with less data. Advances in the underlying algorithms are allowing the neural networks to use data much more efficiently, Carpenter says, enabling training on just a handful of images for some applications. Scientists can also exploit transfer learning, the ability of neural networks to apply classification prowess acquired from one data type to another type. For example, Finkbeiner's team has developed an algorithm that it initially taught to predict cell death on the basis of morphology changes. Although the researchers trained it to study images of rodent cells, it achieved 90% accuracy the first time it was exposed to images of human cells, improving to 99% as it gained experience.

For some of its biological image-recognition work, Google Accelerated Science uses algorithms that were initially trained on hundreds of millions of consumer images mined from the Internet. Researchers then refine that training, using as few as several hundred biological images similar to the ones they wish to study.

Another challenge with deep learning is that the computers are both unintelligent and lazy, notes Michelle Dimon, a research scientist at Google Accelerated Science. They lack the judgement to distinguish biologically relevant differences from normal variation. "The computer is shockingly good at finding batch variation," she notes. As a result, obtaining data that will be fed into a deep-learning algorithm often means applying a high bar for experimental design and controls. Google Accelerated Science requires researchers to place controls randomly on cell-culture plates to account for subtle environmental factors such as incubator temperature, and to use twice as many controls as a biologist might otherwise run. "We make it hard to pipette," Dimon quips.

This hazard underscores the importance of biologists and computer scientists working together to design experiments that incorporate deep learning, Dimon says. And that careful design has become even more important with one of Google's latest projects: Contour, a strategy for clustering cellular-imaging data in ways that highlight trends (such as dose responses) instead of putting them into specific categories (such as alive or dead).

Although deep-learning algorithms can evaluate data without human preconceptions and filters, Greene cautions, that doesn't mean they are unbiased. Training data can be skewed — as happens, for example, when genomic data only from northern Europeans are used. Deep-learning algorithms trained on such data will acquire embedded biases and reflect them in their predictions, which could in turn lead to unequal patient care. If humans help to validate these predictions, that provides a potential check on the problem. But such concerns are troubling if a computer alone is left to make key decisions. "Thinking of these methods as a way to augment humans is better than thinking of these methods as replacing humans," Greene says.

And then there's the challenge of understanding exactly how these algorithms are building the characteristics, or features, that they use to classify data in the first place. Computer scientists are attacking this question by changing or shuffling individual features in a model and then examining how those tweaks change the accuracy of predictions, says Polina Mamoshina, a research scientist at Insilico Medicine in Baltimore, Maryland, which uses deep learning to improve drug discovery. But different neural networks working on the same problem won't approach it in the same way, Greene cautions. Researchers are increasingly focusing on algorithms that make both accurate and explainable predictions, he says, but for now the systems remain black boxes.

"I don't think highly explainable deep-learning models are going to come on the scene in 2018, though I'd love to be wrong," Greene says. ■

Sarah Webb *is a freelance writer in Chattanooga, Tennessee.*

CORRECTION

The Technology feature 'Deep learning for biology' (*Nature* **554,** 555–557; 2018) erroneously affiliated Mark DePristo at Verily Life Sciences. He is, in fact, at Google. Also, the DeepVariant tool was developed jointly by Verily and Google.