MACHINE LEARNING

Do Androids Dream?

The search to understand how artificial neural networks process images yields insights and a trippy brand of beauty

> *Most of the afternoons I would pass looking out at the pasture. I soon began seeing things. A figure emerging from the birch woods and running straight in my direction. Usually it was the Sheep Man, but sometimes it was the Rat, sometimes my girlfriend. Other times it was the sheep with the star on its back.*
>
> —Haruki Murakami, *A Wild Sheep Chase,* 1982

Artificial intelligence has been much in the news lately, driven by ever cheaper computer processing power that has become effectively a near universal commodity. The excitement swirls around mathematical abstractions called deep convolutional neural networks, or ConvNets. Applied to photographs and other images, the algorithms that implement ConvNets identify individuals from their faces, classify objects into one of 1,000 distinct categories (cheetah, husky, strawberry, catamaran, and so on)—and can describe whether they see "two pizzas sitting on top of a stove top oven" or "a red motorcycle parked on the side of the road." All of this happens without human intervention. Researchers looking under the hood of these powerful algorithms are surprised, puzzled and entranced by the beauty of what they find.

BY CHRISTOF KOCH

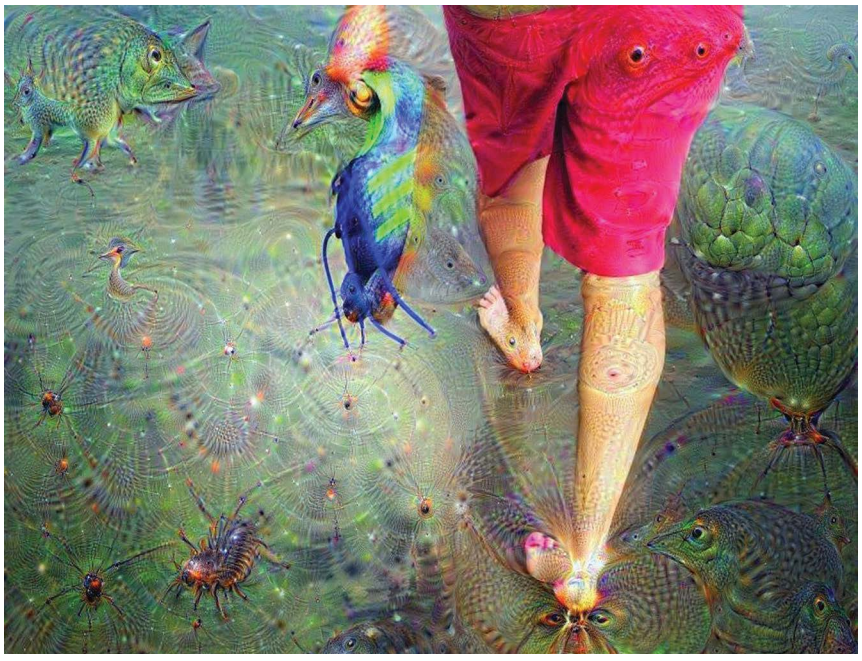
Christof Koch is president and chief scientific officer of the Allen institute for Brain Science in Seattle. He serves on *Scientific American Mind*'s board of advisers.

Springtime for A.I.

How do ConvNets work? Conceptually they are but one or two generations removed from the artificial neural networks developed by engineers and learning theorists in the 1980s and early 1990s. These, in turn, are abstracted from the circuits neuroscientists discovered in the visual system of laboratory animals. Already in the 1950s a few pioneers had found cells in the retinas of frogs that responded vigorously to small, dark spots moving on a stationary background, the famed "bug detectors." Recording from the part of the brain's outer surface that receives visual information, the primary visual cortex, Torsten Wiesel and the late David H. Hubel, both then at Harvard University, found in the early 1960s a set of neurons they called "simple" cells. These neurons responded to a dark or a light bar of a particular orientation in a specific region of the visual field of the animal. Whereas these cells are very particular about where in visual space the oriented line is located, a second set of "complex" cells is less discerning about the exact location of that line. Wiesel and Hubel postulated a wiring scheme to explain their findings, a model that has been enormously influential. It consists of multiple layers of cells—the first layer corresponds to the input cells that carry the visual information as captured by the eyes. These cells respond best to spots of light. They feed into a second layer of neurons, the simple cells, that talk in turn to a third layer of neurons, the complex cells.

Each cell is, in essence, a processing element or unit that computes a weighted sum of its input and, if the sum is sufficiently large, turns the unit's output on; otherwise, it remains off. The exact manner in which the units are wired up determines how cells in the input layer that respond to edges of any orientation are transformed into simple cells that care about a particular orientation and location and then into units that discard some of that spatial information. Subsequent discoveries of neurons in a region of the

LIZ TORMES



A technique called Inceptionism, developed by Google scientists, is used to explore the workings of neural networks. Available as open-source code called DeepDream, it morphs ordinary photographs into bizarrely beautiful images in which eyes, insects and odd creatures emerge from the scene.

monkey visual cortex that switches on for views of faces of monkeys or people reinforced such thinking—visual processing occurs within a hierarchy of processing stages in which the information flows upward, from units that care about low-level features such as brightness, orientation and position to units that represent information in a more abstract manner, such as the presence of any given face or a specific one, such as that of Grandma. Appropriately, this cascade of processing layers is called a feed-forward network.

ConvNets also operate like these specialized networks. A first layer of processing units represents the raw images, whereas subsequent layers extract more and more abstract features. The last output layer may consist of 1,000 units, each representing one of the abovementioned visual-object categories. It effectively decides which object is present in the image. Other signals encode the network's confidence in its final decision.

The modern descendants of these feed-forward networks are bloated, sporting 20 or more layers. Each processing layer has its own wiring scheme, specifying which unit influences which other unit and how strongly it does so. The entire network can have 10 million or more parameters called weights associated with it. And each one must be assigned some numerical value, positive or negative. These legions of numbers cannot be intuited or guessed; they have to be set by hand, an impossible task.

That is where machine learning comes in. Setting these parameters occurs during a learning phase in which the network is shown a million or more pictures of individual objects, together with labels, say, "husky" or "cheetah." Think of Mom showing her toddler a picture book, pointing at a drawing and saying, "Dog." After each such presentation, the network makes a guess based on some initial random setting of its weights.

These are then slightly adjusted to reduce the inevitable mismatch between the output of the network—its guess about what it is looking at—and the correct label. This process repeats over, and over, and over. Supervised learning (the nerdy term is back-propagating the error, or back-prop) is enormously expensive computationally and only became feasible because of the widespread use of so-called graphical processing units developed to support video gaming. Once the training is complete, the network is frozen—it halts the labeling exercises—and can now process novel images, ones it has not previously seen, and can guess their identity, often with near human accuracy.

Machine learning is all the rage in academia and industry, with teams of applied mathematicians and computer scientists competing to develop ever smarter algorithms for optimizing performance.

> CONVOLUTIONAL NETWORKS CAN CORRECTLY IDENTIFY YOUR VACATION PIC AS SHOWING A HUSKY OR A BEGONIA, BUT THEY ALSO **ARRIVE AT NONSENSICAL CONCLUSIONS.**

What Are These Networks Really Doing?

Though relatively simple, ConvNets can yield unexpected surprises. Yes, they can correctly identify your vacation pic as showing a husky or a begonia, but they sometimes also arrive at nonsensical conclusions. A case in point is work by Anh Nguyen and Jeff Clune, both at the University of Wyoming, and Jason Yosinski of Cornell University. To shine light inside the black box of the network, Clune, a computer science professor, and his students developed techniques to dis-

Getting Real

A neural network builds up an image layer by layer.

Processing Layer 1 **Processing Layer 3** **Processing Layer 4**

A deep-learning neural network comes to recognize a tiger by first observing pixels at one network layer that represent lines of color in the animal's fur. It gradually constructs a representation of increasingly abstract features—an eye and later a head—at higher layers of the network.

cover pictures that would evoke strong activation from particular units in a trained ConvNet, asking, "What does this unit really like and want to see?" And how similar would these images be to the pictures that the network encountered during its infancy, when it was being trained? The team started with random images and "evolved" them repeatedly until the network decided, with high confidence, that they were a cheetah, or a handheld remote control, or another visual-object category it had been trained on. The expectation was that the evolutionary algorithm would discover images that most faithfully represented cheetahness, the Platonic idea of a cheetah.

To their surprise, the resultant images were often completely unrecognizable, essentially garbage—colorful, noisy patterns, similar to television static. Although the ConvNet saw, with 99.99 percent confidence, a cheetah in the image, no human would recognize it as a big and very fast African cat. Note that the computer scientists did not modify the ConvNet itself—it still recognized pictures of cheetahs correctly, yet it strangely also insisted that these seemingly noisy images belonged to the same object category. Another way to generate these fooling images yielded pictures that contain bits and pieces of recognizable textures and geometrical structures that the network confidently yet erroneously believed to be a guitar. And these were no rare exceptions.

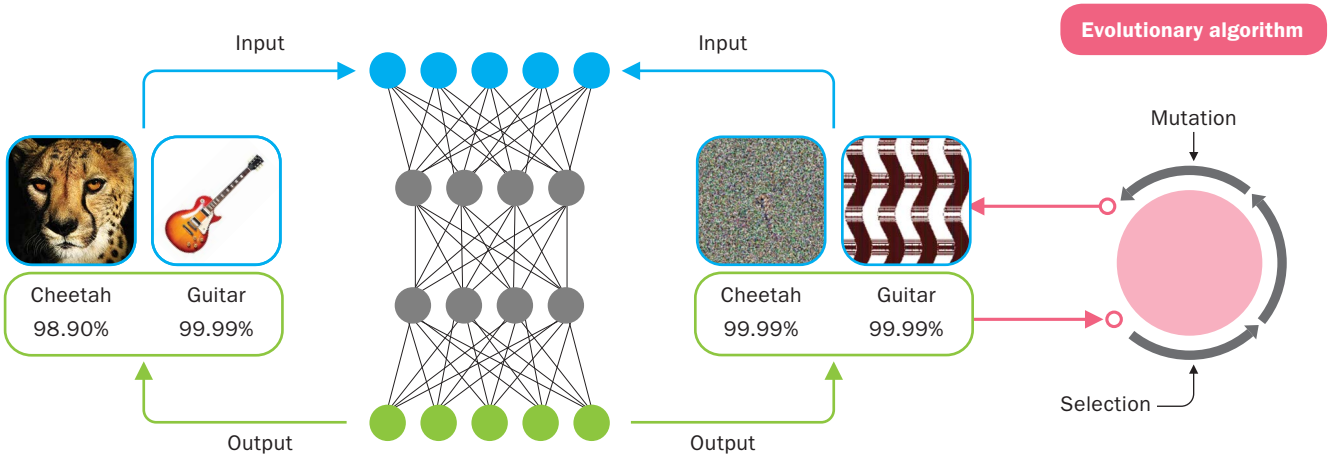
I suspect that if the same image manipulations were to be carried out while recording from the face cells deep inside the visual brain, this procedure would not converge to such nonsensical images but would capture something essential about the nature of faces.

These faux images highlight a large gap between the way people and computers understand visual objects. By watching a cheetah in a zoo or seeing one chase down a gazelle in a nature documentary, we build up an internal representation of these cats that allows us to describe them. If we were forced to, most of us could even draw a cartoon of these graceful animals and specify how they differ from lions or house cats. But computers know nothing of this context. All they have been given are 100 cheetah photographs and a gazillion noncheetah pictures. Without knowing anything about cats—that they have legs, paws, fur, pointed ears, and so on—the network has to figure out what features in the few training images are characteristic of the class of objects known as cheetahs. These same features can also be found in all kinds of other misleading images.

> OVER THE NEXT DECADE A PROPERLY FORMULATED THEORY OF MEANING MAY BRIDGE THE GAP BETWEEN **HOW MACHINES AND HUMANS "SEE" IMAGES.**

This gap had been pointed out for 30 years by American philosopher John Searle in his famous "Chinese room" argument, in which a non-Chinese speaker is able to provide coherent answers to questions posed in Chinese by following a set of carefully laid-out instructions in English to manipulate Chinese characters, even though the individual has no idea what they mean. Searle invented this powerful thought experiment to support his claim that computers, like

GETTY IMAGES



A ConvNet recognizes a cheetah or guitar with a high degree of confidence (*left*), but it can also be fooled if its input image is subject to further processing by an evolutionary algorithm. It might then misidentify a nonsensical image as a cheetah or a guitar (*right*).

SOURCE: "DEEP NEURAL NETWORKS ARE EASILY FOOLED: HIGH CONFIDENCE PREDICTIONS FOR UNRECOGNIZABLE IMAGES," BY ANH NGUYEN ET AL. PRESENTED AT THE 2015 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, BOSTON; JUNE 7–12, 2015 (*neural network* and *patterns*); © ISTOCK.COM (*cheetah* and *guitar*)

the individual in the Chinese room, can never understand anything—they simply follow a set of instructions that makes them appear to be intelligent. That is still true today. But over the next decade machines will become more sophisticated, and it will be more difficult to fool them. The gap between them and us will lessen. Indeed, quite unlike Searle, I do believe that a properly formulated theory of meaning, closely allied to a theory of consciousness, will permit us to ultimately bridge this gap—and truly intelligent machines will then emerge.

Trees Growing Bird Heads

If you are somebody who believes that art and algorithm have nothing but a first letter in common, consider a different way to understand the innards of these networks. In a June 17 blog post, three software engineers, Alexander Mordvintsev, Christopher Olah and Mike Tyka, all at Google, describe a technique termed, in a stroke of marketing genius, Inceptionism, a reference to the popular 2010 psychological science-fiction thriller. The programmers present a starter image to a fully trained machine-learning network and then focus on the artificial neurons in a particular layer between the input layer—equivalent to the retina in an eye—and the final output layer that categorizes an object. The engineers then tweak the input image to maximize the response of the units they are attending to. If they focus on a set of Hubel-and-Wiesel-like units that extract horizontal edges, adding horizontal lines to the original image will enhance their internal response. Or if they focus on units in the upper layers of the network that code for eyes, then inserting eyes into the image will maximize their firing rate. The image is slowly morphed; think of it as controlled hallucination. When focusing on bird units in the upper layers, Inceptionism begins to image birds and superimposes them onto the original images. This turns on the bird units, which further drive the algorithm to enhance the saliency of birds in the image, and so on. Just search online for "Inceptionism," and you will see what I mean. Not surprisingly, the June post has gone viral.

These images are bizarre, strange yet compelling, and often quite pleasing. In an empty sky, birds become visible. Felines are superimposed onto the faces of people in a crowd. A gigantic fish comes to life in the heavens. Patterns imbued with meaning appear in leaves. Castles can be dimly perceived, hovering in the background over an otherwise empty desert landscape. Deep networks go to sleep and dream. It's magical.

Many people have noted the remarkable resemblance between these images and hallucinations produced by tripping on LSD, mescaline or psilocybin mushrooms. In response to the explosion of interest, Google has released open-source code, named DeepDream, to generate such images and assemble them into movies (see http://bit.ly/1FcTca2). For those of us who do not program, a start-up will, for a small fee, modify any image you supply.

For me, as a card-carrying neuroscientist, what is most tantalizing is the remarkable architectural resemblance between the way brains and ConvNets behave. Left to their own devices, what will ConvNets dream about? Electric sheep? Or perhaps a cross between a pig and a snail that shimmers with a psychedelic iridescence? M

MORE TO EXPLORE

- **Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images.** Anh Nguyen et al. Presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston; June 7–12, 2015.
- More examples of images that fooled the computer can be found at **http://evolvingai.org/fooling**