NEWS FEATURE · 09 OCTOBER 2019

Why deep-learning AIs are so easy to fool

Artificial-intelligence researchers are trying to fix the flaws of neural networks.

Douglas Heaven

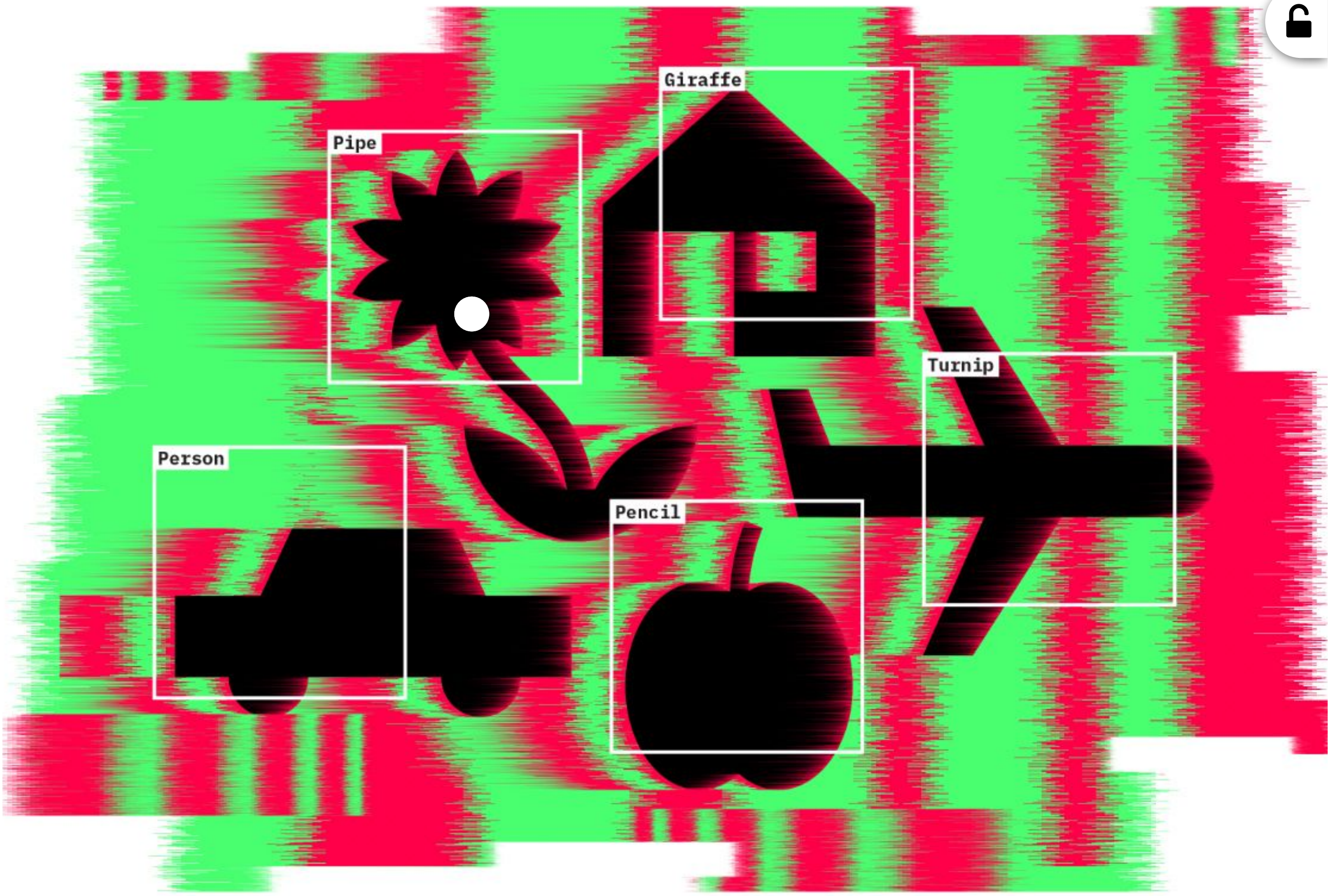

Illustration by Edgar Bąk

A self-driving car approaches a stop sign, but instead of slowing down, it accelerates into the busy intersection. An accident report later reveals that four small rectangles had been stuck to the face the sign. These fooled the car's onboard artificial intelligence (AI) into misreading the word 'stop' as 'speed limit 45'.

Such an event hasn't actually happened, but the potential for sabotaging AI is very real. Researchers have already demonstratedhow to fool an AI system into misreading a stop sign, by carefully positioning stickers on it[1]. They have deceived facial-recognition systemby sticking a printed pattern on glasses or hats. And they havetricked speech-recognition systeminto hearing phantom phrases by inserting patterns of white noise in the audio.

These are just some examples of how easy it is to break the leading pattern-recognition technolog in AI, known as deep neural networks (DNNs). These have proved incredibly successful at correctly classifying all kinds of input, including images, speech and data on consumer preferences. They are part of daily life, running everything from automated telephone systems t user recommendations on the streaming service Netflix. Yet making alterations to inputs — in the form of tiny changes that are typically imperceptible to humans — can flummox the best neural networks around.

These problems are more concerning than idiosyncratic quirks in a not-quite-perfect technolog says Dan Hendrycks, a PhD student in computer science at the University of California, Berkeley. Like many scientists, he has come to see them as the most striking illustration that DNNs are fundamentally brittle: brilliant at what they do until, taken into unfamiliar territory, they break in unpredictable ways.

Sources: Stop sign: Ref. 1; Penguin: Ref. 5

That could lead to substantial problems. Deep-learning systems are increasingly moving out of th lab into the real world, frompiloting self-driving carsto mapping crimeand diagnosing disease But pixels maliciously added to medical scans could fool a DNN into wrongly detecting cancer, one study reported this year[2]. Another suggested that a hacker could use these weaknesses to hijack an online AI-based system so that it runs the invader's own algorithms[3].

In their efforts to work out what's going wrong, researchers have discovered a lot about why

DNNs fail. “There are no fixes for the fundamental brittleness of deep neural networks,” argues François Chollet, an AI engineer at Google in Mountain View, California. To move beyond the flaws, he and others say, researchers need to augment pattern-matching DNNs with extra abilities for instance, making AIs that can explore the world for themselves, write their own code and retai memories. These kinds of system will, some experts think, form the story of the coming decade ir AI research.

Reality check

In 2011, Google revealed a system that could recognize cats in YouTube videos, and soon after came a wave of DNN-based classification systems. “Everybody was saying, ‘Wow, this is amazing, computers are finally able to understand the world,’” says Jeff Clune at the University of Wyoming in Laramie, who is also a senior research manager at Uber AI Labs in San Francisco, California.

But AI researchers knew that DNNs do not actually understand the world. Loosely modelled on the architecture of the brain, they are software structures made up of large numbers of digita neurons arranged in many layers. Each neuron is connected to others in layers above and below it.

The idea is that features of the raw input coming into the bottom layers — such as pixels in an image — trigger some of those neurons, which then pass on a signal to neurons in the layer above according to simple mathematical rules. Training a DNN network involves exposing it to a massive collection of examples, each time tweaking the way in which the neurons are connected so that eventually, the top layer gives the desired answer — such as always interpreting a picture of a lion as a lion, even if the DNN hasn’t seen that picture before.

A first big reality check came in 2013, when Google researcher Christian Szegedy and his colleagues posted a preprint called ‘Intriguing properties of neural networks’[4]. The team showed that it was possible to take an image — of a lion, for example — that a DNN could identify and, by altering a few pixels, convince the machine that it was looking at something different, such as a library. The team called the doctored images ‘adversarial examples’

A year later, Clune and his then-PhD student Anh Nguyen, together with Jason Yosinski at Cornell

University in Ithaca, New York, showed that it was possible to make DNNs see things that were not there, such as a penguin in a pattern of wavy lines[5]. "Anybody who has played with machine learning knows these systems make stupid mistakes once in a while," says Yoshua Bengio at the University of Montreal in Canada, who is a pioneer of deep learning. "What was a surprise was the type of mistake," he says. "That was pretty striking. It's a type of mistake we would not have imagined would happen."

New types of mistake have come thick and fast. Last year, Nguyen, who is now at Auburn University in Alabama, showed that simply rotating objects in an image was sufficient to throw off some of the best image classifiers around[6]. This year, Hendrycks and his colleagues reported that even unadulterated, natural images can still trick state-of-the-art classifiers into making unpredictable gaffes, such as identifying a mushroom as a pretzel or a dragonfly as a manhole cover[7].

The issue goes beyond object recognition: any AI that uses DNNs to classify inputs — such as speech — can be fooled. AIs that play games can be sabotaged: in 2017, computer scientist Sandy Huang, a PhD student at the University of California, Berkeley, and her colleagues focused on DNNs that had been trained to beat Atari video games through a process called reinforcement learning[8]. In this approach, an AI is given a goal and, in response to a range of inputs, learns through trial and error what to do to reach that goal. It is the technology behind superhuman game-playing AIs such as AlphaZero and the poker bot Pluribus. Even so, Huang's team was able to make their AIs lose games by adding one or two random pixels to the screen.

Earlier this year, AI PhD student Adam Gleave at the University of California, Berkeley, and his colleagues demonstrated that it is possible to introduce an agent to an AI's environment that act out an 'adversarial policy' designed to confuse the AI's responses[9]. For example, an AI footballer trained to kick a ball past an AI goalkeeper in a simulated environment loses its ability to score when the goalkeeper starts to behave in unexpected ways, such as collapsing on the ground ɪimulated soccer penalty shootout between two Humanoid robots displayed with and without a versarial policy

An AI footballer in a simulated penalty-shootout is confused when the AI goalkeeper enacts an ‘adversarial policy’: falling to the floor (right). Credit: Adam Gleave/Ref. 9

Knowing where a DNN’s weak spots are could even let a hacker take over a powerful AI. One example of that came last year, when a team from Google showed that it was possible to use adversarial examples not only to force a DNN to make specific mistakes, but also to reprogram it entirely — effectively repurposing an AI trained on one task to do another[3].

Many neural networks, such as those that learn to understand language, can, in principle, be used to encode any other computer program. “In theory, you can turn a chatbot into whatevei programme you want,” says Clune. “This is where the mind starts to boggle.” He imagines a situation in the near future in which hackers could hijack neural nets in the cloud to run their own spambot-dodging algorithms.

For computer scientist Dawn Song at the University of California, Berkeley, DNNs are like sitting ducks. “There are so many different ways that you can attack a system,” she says. “And defence is very, very difficult.”

With great power comes great fragility

DNNs are powerful because their many layers mean they can pick up on patterns in many different features of an input when attempting to classify it. An AI trained to recognize aircraft might find that features such as patches of colour, texture or background are just as stron predictors as the things that we would consider salient, such as wings. But this also means that a very small change in the input can tip it over into what the AI considers an apparently different state.

One answer is simply to throw more data at the AI; in particular, to repeatedly expose the AI to problematic cases and correct its errors. In this form of ‘adversarial training’, as one network learr to identify objects, a second tries to change the first network’s inputs so that it makes mistakes. Ir this way, adversarial examples become part of a DNN’s training data.

Hendrycks and his colleagues have suggested quantifying a DNN's robustness against making errors by testing how it performs against a large range of adversarial examples. However, training a network to withstand one kind of attackcould weaken it against others, they say. And researchers led by Pushmeet Kohli at Google DeepMind in London aretrying to inoculate DNNs against making mistakes Many adversarial attacks work by making tiny tweaks to the component parts of an input — such as subtly altering the colour of pixels in an image — until this tips a DNN over into a misclassification. Kohli's team has suggested that a robust DNN should not change its output as a result of small changes in its input, and that this property might be mathematically incorporated into the network, constraining how it learns

For the moment, however, no one has a fix on the overall problem of brittle AIs. The root of the issue, says Bengio, is that DNNs don't have a good model of how to pick out what matters. When an AI sees a doctored image of a lion as a library, a person still sees a lion because they have a mental model of the animal that rests on a set of high-level features — ears, a tail, a mane and so or — that lets them abstract away from low-level arbitrary or incidental details. "We know from prior experience which features are the salient ones," says Bengio. "And that comes from a deep understanding of the structure of the world."

One attempt to address this is to combine DNNs with symbolic AI, which was the dominant paradigm in AI before machine learning. With symbolic AI, machines reasoned using hard-coded rules about how the world worked, such as that it contains discrete objects and that they are related to one another in various ways. Some researchers, such as psychologist Gary Marcus at New York University, say hybrid AI models are the way forward. "Deep learning is so useful in the short term that people have lost sight of the long term," says Marcus, who is a long-time critic of the current deep-learning approach. In May, he co-founded a start-up called Robust AI in Palo Alto, California, which aims to mix deep learning with rule-based AI techniques to develop robots that can operate safely alongside people. Exactly what the company is working on remains under wraps.

Even if rules can be embedded into DNNs, they are still only as good as the data they learn from. Bengio says that AI agents need to learn in richer environments that they can explore. For exampl most computer-vision systems fail to recognize that a can of beer is cylindrical because they wer trained on data sets of 2D images. That is why Nguyen and colleagues found it so easy to fool DNNs by presenting familiar objects from different perspectives. Learning in a 3D environment —

real or simulated — will help.

But the way AIs do their learning also needs to change. “Learning about causality needs to be done by agents that do things in the world, that can experiment and explore,” says Bengio. Another deep-learning pioneer, Jürgen Schmidhuber at the Dalle Molle Institute for Artificial Intelligence Research in Manno, Switzerland, thinks along similar lines. Pattern recognition is extremely powerful, he says — good enough to have made companies such as Alibaba, Tencent, Amazon, Facebook and Google the most valuable in the world. “But there’s a much bigger wave coming,” he says. “And this will be about machines that manipulate the world and create their own data through their own actions.”

In a sense, AIs that use reinforcement learning to beat computer games are doing this already in artificial environments: by trial and error, they manipulate pixels on screen in allowed ways until they reach a goal. But real environments are much richer than the simulated or curated data sets c which most DNNs train today.

Robots that improvise

In a laboratory at the University of California, Berkeley, a robot arm rummages through clutter. It picks up a red bowl and uses it to nudge a blue oven glove a couple of centimetres to the right. It drops the bowl and picks up an empty plastic spray bottle. Then it explores the heft and shape of a paperback book. Over several days of non-stop sifting, the robot starts to get a feel for these alien objects and what it can do with them.

The robot arm is using deep learning to teach itself to use tools. Given a tray of objects, it picks up and looks at each in turn, seeing what happens when it moves them around and knocks one object into another.

llection of clips of robots improvising with tools

Robots use deep learning to explore how to use 3D tools. Credit: Annie Xie

When researchers give the robot a goal — for instance, presenting it with an image of a nearly empty tray and specifying that the robot arrange objects to match that state — it improvises, and can work with objects it has not seen before, such as using a sponge to wipe objects off a table. It also figured out that clearing up using a plastic water bottle to knock objects out of the way is quicker than picking up those objects directly. “Compared to other machine-learning technique the generality of what it can accomplish continues to impress me,” says Chelsea Finn, who worked at the Berkeley lab and is now continuing that research at Stanford University in California.

This kind of learning gives an AI a much richer understanding of objects and the world in general, says Finn. If you had seen a water bottle or a sponge only in photographs, you might be able to recognize them in other images. But you would not really understand what they were or what they could be used for. “Your understanding of the world would be much shallower than if you could actually interact with them,” she says.

But this learning is a slow process. In a simulated environment, an AI can rattle through examples at lightning speed. In 2017, AlphaZero, the latest version of DeepMind’s self-taught game-playing software, was trained to become a superhuman player of Go, then chess and then shogi (a form of Japanese chess) in just over a day. In that time, it played more than 20 million training games of each event.

AI robots can’t learn this quickly. Almost all major results in deep learning have relied heavily on large amounts of data, says Jeff Mahler, co-founder of Ambidextrous, an AI and robotics company in Berkeley, California. “Collecting tens of millions of data points would cost years of continuous execution time on a single robot.” What’s more, the data might not be reliable, because the calibration of sensors can change over time and hardware can degrade

Because of this, most robotics work that involves deep learning still uses simulated environments to speed up the training. “What you can learn depends on how good the simulators are,” says David Kent, a PhD student in robotics at the Georgia Institute of Technology in Atlanta. Simulators are improving all the time, and researchers are getting better at transferring lessons learnt in virtu worlds over to the real. Such simulations are still no match for real-world complexities, however

Finn argues that learning using robots is ultimately easier to scale up than learning with artificial data. Her tool-using robot took a few days to learn a relatively simple task, but it did not require heavy monitoring. “You just run the robot and just kind of check in with it every once in a while,” she says. She imagines one day having lots of robots out in the world left to their own devices, learning around the clock. This should be possible — after all, this is how people gain an understanding of the world. “A baby doesn’t learn by downloading data from Facebook,” says Schmidhuber.

Learning from less data

A baby can also recognize new examples from just a few data points: even if they have never seen a giraffe before, they can still learn to spot one after seeing it once or twice. Part of the reason thi: works so quickly is because the baby has seen many other living things, if not giraffes, so is alread familiar with their salient features.

A catch-all term for granting these kinds of abilities to AIs is transfer learning: the idea being to transfer the knowledge gained from previous rounds of training to another task. One way to do this is to reuse all or part of a pre-trained network as the starting point when training for a new task. For example, reusing parts of a DNN that has already been trained to identify one type of animal — such as those layers that recognize basic body shape — could give a new network the edge when learning to identify a giraffe.

An extreme form of transfer learning aims to train a new network by showing it just a handful of examples, and sometimes only one. Known as one-shot or few-shot learning, this relies heavily on pre-trained DNNs. Imagine you want to build a facial-recognition system that identifies people in a criminal database. A quick way is to use a DNN that has already seen millions of faces (not necessarily those in the database) so that it has a good idea of salient features, such as the shapes noses and jaws. Now, when the network looks at just one instance of a new face, it can extract a useful feature set from that image. It can then compare how similar that feature set is to those o single images in the criminal database, and find the closest match

Having a pre-trained memory of this kind can help AIs to recognize new examples without needing to see lots of patterns, which could speed up learning with robots. But such DNNs might still be at a loss when confronted with anything too far from their experience. It’s still not clear ho much these networks can generalize.

Even the most successful AI systems such as DeepMind’s AlphaZero have an extremely narrow sphere of expertise. AlphaZero’s algorithm can be trained to play both Go and chess, but not both at once. Retraining a model’s connections and responses so that it can win at chess resets an previous experience it had of Go. “If you think about it from the perspective of a human, this is kind of ridiculous,” says Finn. People don’t forget what they’ve learnt so easily.

Learning how to learn

AlphaZero’s success at playing games wasn’t just down to effective reinforcement learning, but also to an algorithm that helped it (using a variant of a technique called Monte Carlo tree search) t narrow down its choices from the possible next steps[10]. In other words, the AI was guided in how best to learn from its environment. Chollet thinks that an important next step in AI will be to give DNNs the ability to write their own such algorithms, rather than using code provided by humans.

Supplementing basic pattern-matching with reasoning abilities would make AIs better at dealing with inputs beyond their comfort zone, he argues. Computer scientists have for years studiec program synthesis, in which a computer generates code automatically. Combining that field witl deep learning could lead to systems with DNNs that are much closer to the abstract mental model that humans use, Chollet thinks.

In robotics, for instance, computer scientist Kristen Grauman at Facebook AI Research in Menlc Park, California, and the University of Texas at Austin is teaching robots how best to explore new environments for themselves. This can involve picking in which directions to look when presentec with new scenes, for instance, and which way to manipulate an object to best understand its shape or purpose. The idea is to get the AI to predict which new viewpoint or angle will give it the most useful new data to learn from.

Researchers in the field say they are making progress in fixing deep learning’s flaws, but acknowledge that they’re still groping for new techniques to make the process less brittle. There i not much theory behind deep learning, says Song. “If something doesn’t work, it’s difficult to figure out why,” she says. “The whole field is still very empirical. You just have to try things.”

For the moment, although scientists recognize the brittleness of DNNs and their reliance on larg amounts of data, most say that the technique is here to stay. The realization this decade that neur networks — allied with enormous computing resources — can be trained to recognize patterns so well remains a revelation. “No one really has any idea how to better it,” says Clune.

Nature **574**, 163-166 (2019)

doi: 10.1038/d41586-019-03013-5

References

1. Eykholt, K. *et al. IEEE/CVF Conf. Comp. Vision Pattern Recog.* **2018**, 1625–1634 (2018).
2. Finlayson, S. G. *et al. Science* **363**, 1287–1289 (2019).

show more ⌄

Nature ISSN 1476-4687 (online)