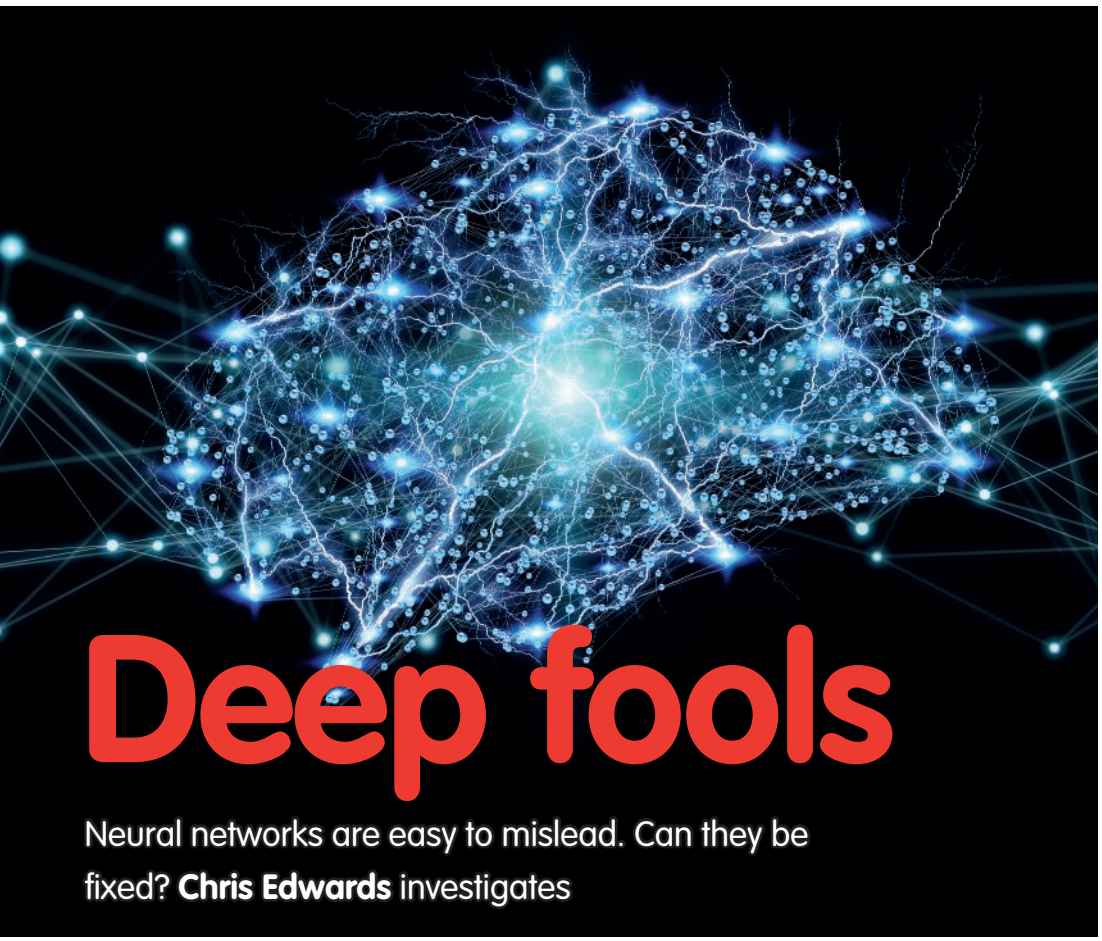


Deep fools

Neural networks are easy to mislead. Can they be fixed? **Chris Edwards** investigates

In not much more than a decade, deep learning has moved from a research curiosity to a technology that could underpin a new generation of autonomous vehicles and robots: machines that can respond more intelligently to the world they perceive.

Among the examples of how deep learning can work in these systems was a 2011 project at the Swiss research institute IDSIA. The GPU-assisted deep neural network (DNN) could correctly read signs that, to the naked eye, looked completely bleached out. Six years later, University of Michigan PhD student Kevin Eykholt and colleagues put brightly coloured stickers on normal road signs and in doing so fooled a DNN into thinking a stop sign said "45mph speed limit". A later experiment rendered the stop sign practically invisible to the computer.

The results might have surprised DNN researchers had they not followed a similar profile to those discovered by Christian Szegedy and colleagues at Google in 2013 when they discovered many tiny changes to an image would lead to bizarrely wrong misclassifications. Schools buses turned into birds – but to the naked eye there was no difference in correct and misclassified images.

In a keynote speech at the 2018 Deep Learning and Security workshop Google Brain research scientist Ian Goodfellow said of Szegedy, "He wasn't trying to break the network, he was trying to analyse it. He thought if you tried to change a schoolbus into an ostrich you would make it grow feathers."

Szegedy's work uncovered a key aspect of the way in which DNNs analyse images. They often home in on tiny features that are imperceptible to the human eye. And, as later work by Anh Nguyen, assistant professor at Auburn University, has shown they do not readily separate objects within an image the same way biological brains do. Trained on images of weightlifters, the DNN will readily treat an arm and dumbell as one discrete object.

Although a lot of the demonstrations of the fallibility of deep learning have focused on image manipulation, a number of the researchers investigating how to make AI more robust started with an interest in security and malware detection. One of the earliest applications for AI-like software was in email spam filtering and a major thrust in development now is in intrusion detection that can move beyond simple rules.

Now working at Google, Nicholas Carlini worked on malware detection for his PhD dissertation at the University of California at Berkeley but while taking a break after writing up his research on AI security found a way to break the DeepSpeech voice recognition system published by Mozilla in a matter of hours. "The reason why it took less than two days was because I had been doing this work, looked at the system and thought 'let's put the two together'," he says.

Though Carlini was not able to make the attack work with audio picked up a voice system's microphone but only on digital files supplied to DeepSpeech directly, a series of tiny manipulations that would only be detected as noise by a human observer made it possible to turn a voice sample into any command Carlini wanted or even to convince the system it heard nothing. Similar attacks by other teams have succeeded with over-the-air attacks, though they have not so far been able to substitute arbitrary phrases for each other.

Cat-and-mouse

The cat-and-mouse game discovering and fixing vulnerabilities among software that ranges from voice

response to network intrusion detection illustrates a problem that many AI systems that take commands from a variety of sources tend to face.

The potential targets range from factory-floor cobots to autonomous vehicles, which not only will take image inputs but will have to deal with a variety of sensor inputs while hackers attempt attacks over a network. Would confetti scattered in the path of oncoming cars make them unable to detect stop lights and pedestrians crossing a street? Are there attacks on other sensors that might prevent a robot from stopping before it injures someone or damages itself?

In their report 'The Malicious Use of Artificial Intelligence' published early last year, a group of authors from 14 institutions around the world concluded: "We should expect attacks that exploit the vulnerabilities of AI systems to become more typical. This prediction follows directly from the unresolved vulnerabilities of AI systems and the likelihood that AI systems will become increasingly pervasive."

The problem that faces defenders against the dark arts of AI manipulation is that why they work is not well understood. Numerous research teams, mostly working on images, have come up with proposals for dealing with the adversarial examples that upset DNNs. But, as Goodfellow points out: "Quite a lot of those papers get broken."

A key problem for DNNs lies in their very depth. A typical network can contain millions of parameters. The huge number of dimensions in the model, tend to soften the distinctions between seemingly quite distant classifications. Mainuddin Jonas, a PhD student working in Professor David Evans' group at the University of Virginia published work last year that shows how it is possible to guide a DNN away from the correct classification layer by layer through numerous small tweaks that push neuronal outputs away from the correct targets. Random noise does not have the same effect.

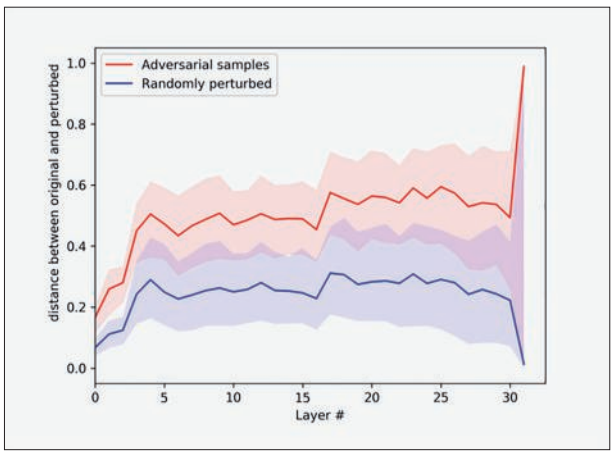
Manipulation

DNNs are not alone in being prone to this kind of manipulation. Even before Szegedy discovered the problems with DNNs, Battista Biggio at the University of Cagliari working with researchers from the University of Tubingen found support vector machines are vulnerable to adversarial examples. Spammers have also taken advantage of similar approaches to defeat filters that use Bayesian classification. Research at Google showed that humans can also be fooled in similar ways if they have only a fraction of a second to view an image.

For those designing robotic systems that use deep learning to provide better information for their control systems, another stumbling block besides high dimensionality is determining ground truth for inputs from the real world. Evans says: "Natural images and audio are inherently ambiguous. The question comes down to how a human will interpret something."

In robotic control, ideally the system designers want something that outperforms humans: recognising signs of danger long before we would pick it up. Some comfort may come from research on malware detection. Carlini says in principle it is easier to determine ground truth for software behaviour. "I can list the things that I want the program to do. That can be checked. Did it create this file? Yes or no. There is no human perception there. Though it's somewhat complicated to say what is and what's not malware, we can always start by defining what is malicious behaviour. And we can find adversarial example that say 'this is benign' but is in fact malicious."

"In not much more than a decade, deep learning has moved from a research curiosity to a technology that could underpin a new generation of machines that can respond intelligently to the world around them."



Below: This figure shows how artificially and randomly altered images move through the successive layers of a neural network.

As a result, malware classification provides an appealing domain for AI-security research. But will the experience in domains that offer solid ground truths cross over into areas such as image recognition? "What we've learned so far about adversarial malware tends to be specific to the type of malware," Carlini says.

One way to reduce the ability of hackers to find adversarial examples is to constrain the inputs to the neurons, an approach Evans calls feature squeezing, and so reduce the hacker's search space for successful adversarial examples. Reworking the input is an technique that Nguyen and colleagues have used. They converted pixel images of handwritten characters to vectorised forms that "purify" the image and remove much of the artificially induced noise.

Another option with images and sound is to enlist the help of a different kind of neural network: a deep generation network (DGN), a technology that became famous for synthesising believable faces and realistic landscapes from minimal information. As a sanity check, the DGN would reconstruct what the network used to recognise and classify images claims it is seeing. "Because the DGN is trained to only generate realistic-looking images, the matched, generated image would look real," Nguyen says.

In doing so, robots that attempt to process the world as it is may have to synthesise their own models of the world just to check they are not being fooled.