Optical illusions fool computers into seeing things

16:10 11 December 2014 by **Jacob Aron**

Computers are starting to identify objects with near-human levels of accuracy, enabling them to do everything from creating automatic picture captions to driving cars. But now a collection of bizarre optical illusions for these artificial-intelligence systems (AIs) has revealed that machines don't see the same way we do, which could leave them vulnerable to exploitation.

Image-recognition algorithms learn to recognise objects by training on a large number of images and identifying patterns that mark out a cat from a coffee cup, for example.

Jeff Clune of the University of Wyoming in Laramie and his colleagues wanted to know if they could hook up a particular type of image-recognition algorithm called a deep neural network (DNN) to a second algorithm designed to evolve different pictures.

Such genetic algorithms, working in conjunction with human judgement, have previously created images of apples and faces, so Clune wondered if replacing the human with a DNN, to work alongside the genetic algorithm, would work as well, resulting in a computer that could generate creative pictures by itself.

"We were expecting that we would get the same thing, a lot of very high-quality recognisable images," Clune says. "Instead we got these rather bizarre images: a cheetah that looks nothing like a cheetah."

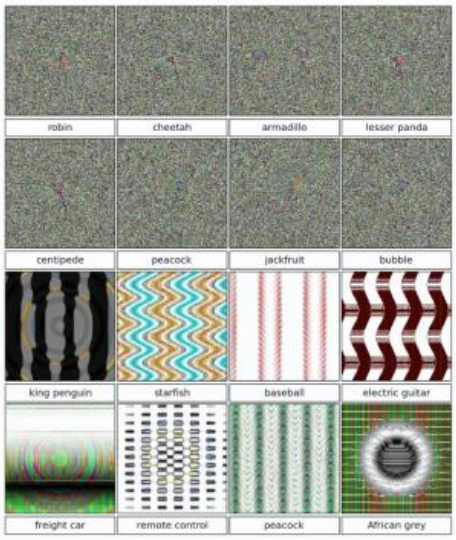
Easily fooled

Clune used one of the best DNNs, called AlexNet, created by researchers at the University of Toronto, Canada, in 2012 – its performance is so impressive that Google hired them last year.

It turned out that the genetic algorithm produced images of seemingly random static that AlexNet declared to be pictures of a variety of animals with more than 99 per cent certainty (see picture). Other images, generated in a different way, look like vaguely evocative abstract art to humans, but fool AlexNet into seeing a baseball, electric guitar or other household object.

The algorithm's confusion is due to differences in how it sees the world compared with humans, says Clune. While we identify a cheetah by looking for the whole package – the right body shape, patterning and so on – a DNN is only interested in the parts of an object that most distinguish it from others. "It's almost like these DNNs are huge fans of cubist art," says Clune.

Optical illusions



The algorithm was nearly 100 per cent convinced it had labelled these images correctly *(Image: Jason Yosinski, Jeff Clune and Anh Nguyen)*

In a way, it's not surprising that image-recognition algorithms can be fooled – after all, optical illusions catch us humans out all the time. "All optical illusions are kind of hacking the human visual system, in the same sense that our paper is hacking the DNN visual system to fool it into seeing something that isn't there," says Clune.

Studying these illusions and the differences between algorithms and humans could teach us more about ourselves, says Jürgen Schmidhuber of the Dalle Molle Institute for Artificial Intelligence Research in Manno, Switzerland. "These networks make predictions about what neuroscientists will find in a couple of decades, once they are able to decode and ready the synapses in the human brain."

More immediately, Clune wants to figure out how to help DNNs ignore the illusions. If they can be fooled by static, an attacker may be able to bypass facial-recognition security systems, or even trick driverless cars into seeing misleading road signs. "It opens any application that uses this computer vision up to security hacks," he says.

Journal reference: arxiv.org/abs/1412.1897

More Latest news

Try the hardest crossword ever set by a computer


18:00 27 December 2014
Silicon-chip logic is remorseless, but it can think laterally enough to flummox human minds. Up for the challenge? There's a prize to be won if you are

3D wall immerses you in an earthquake