When all the world's a toaster, according to tricked AI

January 5 2018, by Nancy Owano

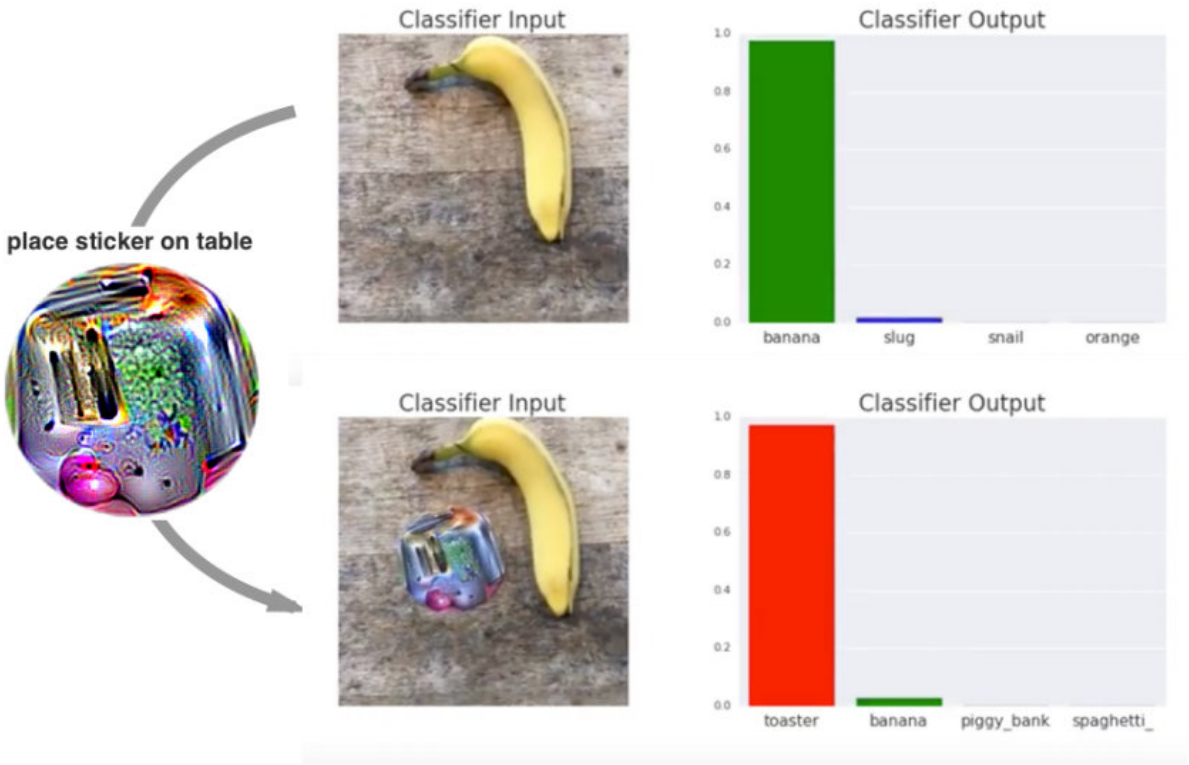

Image recognition technology can be duped by psychedelic stickers created by a Google team.

The stickers made the tech to "see" things that were not there.

Emma Sims in *Alphr*: "machine learning systems can be distracted with highly localised psychedelic stickers, causing an oversight in broader computer vision."

The lowly toaster decked out with computer generated patterns took center-stage as a research example, with the toaster-inspired patterns distracting image recognition software. In short, Google researchers figured out how to trick neural networks into thinking something not a toaster was a toaster.

"AI loves a psychedelic aesthetic." That was a subhead in *Alphr* and that was exactly the problem that pushed a machine learning system off the rails.

Think about that. Sims mused, "In other words, something that would sell like hotcakes for £8.99 in Urban Outfitters has the capacity to deceive highly advanced machine learning systems."

As described in BBC News: "When the patterns were put next to another item, such as a banana, many neural networks saw a toaster instead."

Sims in *Alphr* explained what made the stickers so intoxicating. "AI uses cognitive shortcuts, as humans do, to visually apprehend images." The Google team came up with mesmerising visuals on which AI involuntarily fixated; the "funky psychedelic stickers" lured the AI "away from what it should be focusing on."

Google researcher Tom Brown said, "Our adversarial patch is more effective than a picture of a real toaster, even when the patch is significantly smaller than the toaster."

BBC News quoted them as saying, "These adversarial patches can be printed, added to any scene, photographed, and presented to image

classifiers."

The team wrote a paper discussing their work, titled "Adversarial Patch." The patch is described in detail. Authors are Tom Brown, Dandelion Mané, Aurko Roy, Martín Abadi and Justin Gilmer; the paper is on arXiv.

When a photo of a tabletop with a banana and a notebook is passed through VGG16, the team said in their paper, the network reports class 'banana' with 97% confidence. If they place a sticker targeted to the class "toaster" on the table, the photograph is classified as a toaster with 99% confidence.

What is VGG16? This is a convolutional neural network architecture named after the Visual Geometry Group from Oxford, who developed it.

The BBC report noted that the pattern consistently tricked image recognition software when it took up at least 10% of a scene.

Because this patch is scene-independent, it allows attackers to create a physical-world attack without prior knowledge of the lighting conditions, camera angle, type of classifier being attacked, or even the other items within the scene, the authors said.

Thomas Claburn in *The Register* raised the point that "The attack differs from other approaches in that it doesn't rely on altering an image with graphic artifacts. Rather, it involves adding the adversarial patch to the scene being captured by image recognition software."

Google's researchers, as part of their continued exploration into artificial intelligence, are always interested in learning how AI might be tricked. "The team said the method could be used to 'attack' image recognition systems," said the BBC.

"Practical applications for the discovery include the bypassing of security systems at airports or prisons, allowing contraband material to elude recognition," said Sims.

More information: Adversarial Patch, arXiv:1712.09665 [cs.CV] arxiv.org/abs/1712.09665

Abstract

We present a method to create universal, robust, targeted adversarial image patches in the real world. The patches are universal because they can be used to attack any scene, robust because they work under a wide variety of transformations, and targeted because they can cause a classifier to output any target class. These adversarial patches can be printed, added to any scene, photographed, and presented to image classifiers; even when the patches are small, they cause the classifiers to ignore the other items in the scene and report a chosen target class.