These stickers make computer vision software hallucinate things that aren’t there

If you’re trying to fool an AI vision system in the future, then stickers like these might help.

For a long time, computer scientists have been developing special types of images that bamboozle AI eyes. These pictures and patterns are known as “adversarial images,” and they exploit weaknesses in the way computers look at the world to make them see stuff that isn’t there. Think of them as optical illusions, but for AI. They can be made into glasses that fool facial recognition systems, they can be printed onto physical objects, and now, researchers from Google have turned them into stickers.

A paper describing the production of these stickers was published last month, and although the work isn’t a breakthrough, it is a neat step forward. As well as the fact that these adversarial images can be printed at home (you can even do so yourself if you like), they’re also remarkably flexible. Unlike other adversarial attacks, they don’t need to be tuned based on the image they’re trying to override, nor does it matter where they appear in the AI’s field of view. Here’s what it looks like in action, with a sticker that turns a banana into a toaster:

[embedded content]

As the researchers write, the sticker “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.” So, after such an image is generated, it could be “distributed across the Internet for other attackers to print out and use.”

This is why many AI researchers are worried about how these methods might be used to attack systems like self-driving cars. Imagine a little patch you can stick onto the side of the motorway that makes your sedan think it sees a stop sign, or a sticker that stops you from being identified up by AI surveillance systems. “Even if humans are able to notice these patches, they may not understand the intent [and] instead view it as a form of art,” the researchers write.

There’s no need to worry about such attacks yet, though. Although adversarial images can be disconcertingly effective, they’re not some super magic hack that works on every AI system every time. Patches like the one the Google researchers created take time and effort to generate, and usually access to the code of the vision systems they’re targeting. The problem, as research like this shows, is that these attacks are slowly getting more flexible and effective over time. Stickers might just be the start.

https://www.theverge.com/2018/1/3/16844842/ai-computer-vision-trick-adversarial-patches-google