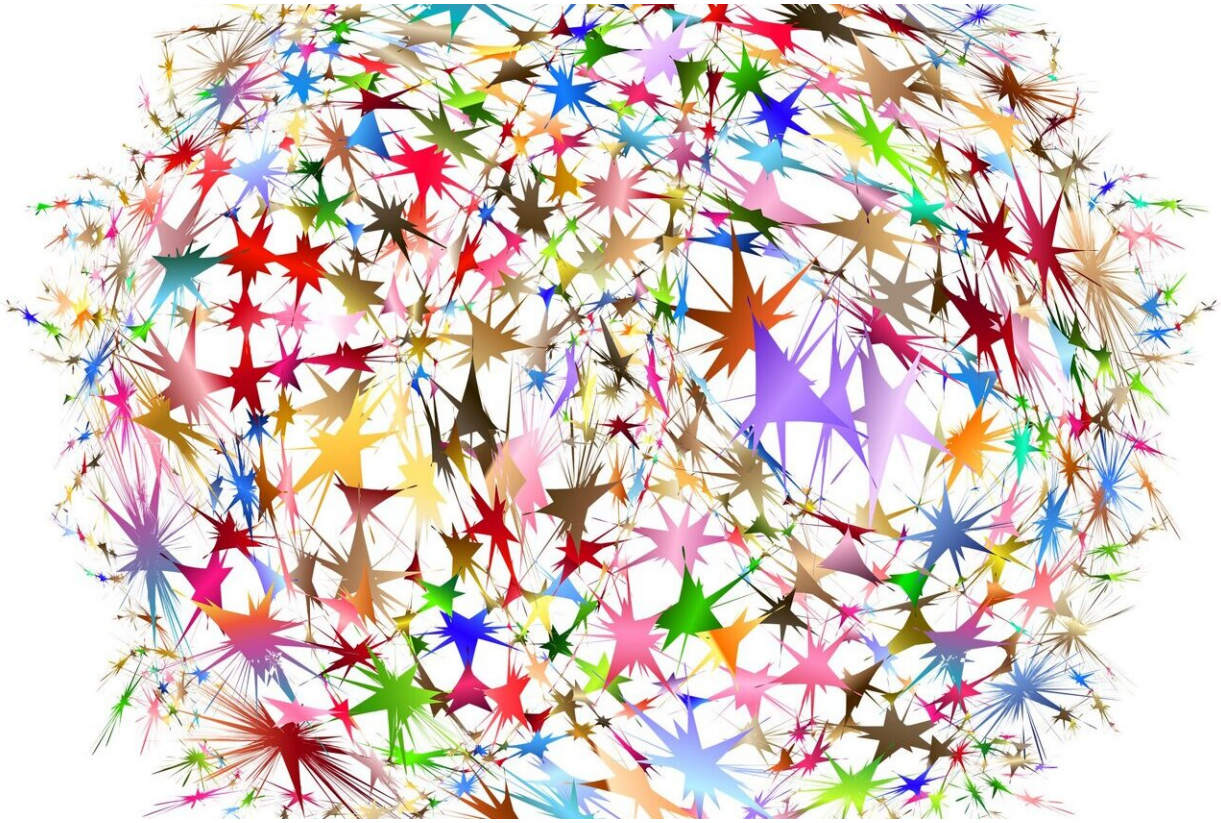
Why did my classifier just mistake a turtle for a rifle?

August 1 2019, by Kim Martineau



Credit: CC0 Public Domain

A few years ago, the idea of tricking a computer vision system by subtly altering pixels in an image or hacking a street sign seemed like more of a hypothetical threat than anything to seriously worry about. After all, a

self-driving car in the real world would perceive a manipulated object from multiple viewpoints, cancelling out any misleading information. At least, that's what one study claimed.

"We thought, there's no way that's true!" says MIT Ph.D. student Andrew Ilyas, then a sophomore at MIT. He and his friends—Anish Athalye, Logan Engstrom, and Jessy Lin—holed up at the MIT Student Center and came up with an experiment to refute the study. They would print a set of three-dimensional turtles and show that a computer vision classifier could mistake them for rifles.

The results of their experiments, published at last year's International Conference on Machine Learning (ICML), were widely covered in the media, and served as a reminder of just how vulnerable the artificial intelligence systems behind self-driving cars and face-recognition software could be. "Even if you don't think a mean attacker is going to perturb your stop sign, it's troubling that it's a possibility," says Ilyas. "Adversarial example research is about optimizing for the worst case instead of the average case."

With no faculty co-authors to vouch for them, Ilyas and his friends published their study under the pseudonym "Lab 6," a play on Course 6, their Department of Electrical Engineering and Computer Science (EECS) major. Ilyas and Engstrom, now an MIT graduate student, would go on to publish five more papers together, with a half-dozen more in the pipeline.

At the time, the risk posed by adversarial examples was still poorly understood. Yann LeCun, the head of Facebook AI, famously downplayed the problem on Twitter. "Here's one of the pioneers of deep learning saying, this is how it is, and they say, nah!" says EECS Professor Aleksander Madry. "It just didn't sound right to them and they were determined to prove why. Their audacity is very MIT."

The extent of the problem has grown clearer. In 2017, IBM researcher Pin-Yu Chen showed that a computer vision model could be compromised in a so-called black-box attack by simply feeding it progressively altered images until one caused the system to fail. Expanding on Chen's work at ICML last year, the Lab 6 team highlighted multiple cases in which classifiers could be duped into confusing cats and skiers for guacamole and dogs, respectively.

This spring, Ilyas, Engstrom, and Madry presented a framework at ICML for making black-box attacks several times faster by exploiting information gained from each spoofing attempt. The ability to mount more efficient black-box attacks allows engineers to redesign their models to be that much more resilient.

"When I met Andrew and Logan as undergraduates, they already seemed like experienced researchers," says Chen, who now works with them via the MIT-IBM Watson AI Lab. "They're also great collaborators. If one is talking, the other jumps in and finishes his thought."

That dynamic was on display recently as Ilyas and Engstrom sat down in Stata to discuss their work. Ilyas seemed introspective and cautious, Engstrom, outgoing, and at times, brash.

"In research, we argue a lot," says Ilyas. "If you're too similar you reinforce each other's bad ideas." Engstrom nodded. "It can get very tense."

When it comes time to write papers, they take turns at the keyboard. "If it's me, I add words," says Ilyas. "If it's me, I cut words," says Engstrom.

Engstrom joined Madry's lab for a SuperUROP project as a junior; Ilyas joined last fall as a first-year Ph.D. student after finishing his undergraduate and MEng degrees early. Faced with offers from other

top graduate schools, Ilyas opted to stay at MIT. A year later, Engstrom followed.

This spring the pair was back in the news again, with a new way of looking at adversarial examples: not as bugs, but as features corresponding to patterns too subtle for humans to perceive that are still useful to learning algorithms. We know instinctively that people and machines see the world differently, but the paper showed that the difference could be isolated and measured.

They trained a model to identify cats based on "robust" features recognizable to humans, and "non-robust" features that humans typically overlook, and found that visual classifiers could just as easily identify a cat from non-robust features as robust. If anything, the model seemed to rely more on the non-robust features, suggesting that as accuracy improves, the model may become more susceptible to adversarial examples.

"The only thing that makes these features special is that we as humans are not sensitive to them," Ilyas told Wired.

Their eureka moment came late one night in Madry's lab, as they often do, following hours of talking. "Conversation is the most powerful tool for scientific discovery," Madry likes to say. The team quickly sketched out experiments to test their idea.

"There are many beautiful theories proposed in deep learning," says Madry. "But no hypothesis can be accepted until you come up with a way of verifying it."

"This is a new field," he adds. "We don't know the answers to the questions, and I would argue we don't even know the right questions. Andrew and Logan have the brilliance and drive to help lead the way."

More information: NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles, arXiv:1707.03501 [cs.CV] arxiv.org/abs/1707.03501

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology