Slight Street Sign Modifications Can Completely Fool Machine Learning Algo...

Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms

Minor changes to street sign graphics can fool machine learning algorithms into thinking the signs say something completely different

BY EVAN ACKERMAN

04 AUG 2017 | 5 MIN READ |

PHOTO: CORNELL UNIVERSITY

AUTOMOTIVE SENSORS | SENSORS | NEURAL NETWORKS | CLASSIFICATION | AUTONOMOUS VEHICLES

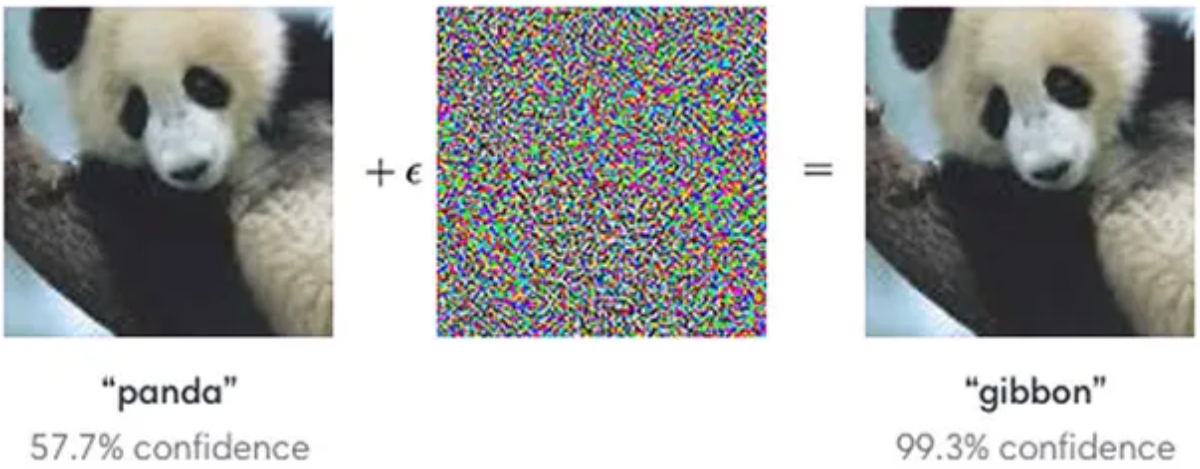
It's very difficult, if not impossible, for us humans to understand how robots see the world. Their cameras work like our eyes do, but the space between the image that a camera captures and actionable information about that image is filled with a black box of machine learning algorithms that are trying to translate patterns of features into something that they're familiar with. Training these algorithms usually involves showing them a set of different pictures of something (like a stop sign), and then seeing if they can extract enough common features from those pictures to reliably identify stop signs that aren't in their training set.

This works pretty well, but the common features that machine learning algorithms come up with generally are not "red octagons with the letters S-T-O-P on them." Rather, they're looking features that all stop signs share, but would not be in the least bit comprehensible to a human looking at them. If this seems hard to visualize, that's because it reflects a fundamental disconnect between the way our brains and artificial neural networks interpret the world.

The upshot here is that slight alterations to an image that are invisible to humans can result in wildly different (and sometimes bizarre) interpretations from a machine learning algorithm. These "adversarial images" have generally required relatively complex analysis and image manipulation, but a group of researchers from the University of

the University of California Berkeley have just published a paper showing that it's also possible to trick visual classification algorithms by making slight alterations in the physical world. A little bit of spray paint or some stickers on a stop sign were able to fool a deep neural network-based classifier into thinking it was looking at a speed limit sign 100 percent of the time.

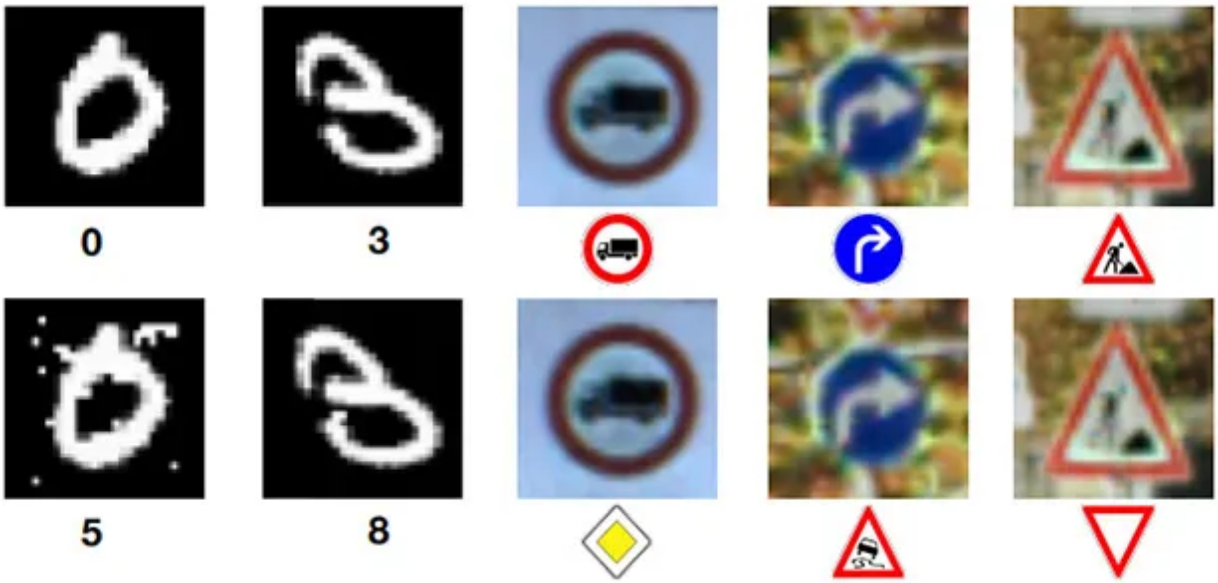
Here's an example of the kind of adversarial image we're used to seeing:



An image of a panda, when combined with an adversarial input, can convince a classifier that it's looking at a gibbon. IMAGE: OPENAI

Obviously, it's totally, uh, obvious to us that both images feature a panda. The differences between the first and third images are invisible to us, and even when the alterations are shown explicitly, there's nothing in there that looks all that much like a gibbon. But to a neural

third image is almost definitely a gibbon. This kind of thing also works with street signs, causing signs that look like one thing to us to look like something completely different to the vision system of an autonomous car, which could be very dangerous for obvious reasons.



Top row shows legitimate sample images, while the bottom row shows adversarial sample images, along with the output of a deep neural network classifier below each image. IMAGES: PAPERNOT ET AL

Adversarial attacks like these, while effective, are much harder to do in practice, because you usually don't have direct digital access to the inputs of the neural network you're trying to mess with. Also, in the context of something like an autonomous car, the neural network has the opportunity to analyze a whole bunch of images of a sign at different distances and angles as it approaches. And lastly, adversarial images tend to include introduced features over the entire image (both the sign and the background), which doesn't work in real life.

What's novel about this new technique is that it's based on *physical*

a way that they reliably screw up neural network classifiers from multiple distances and angles while remaining discreet enough to be undetectable to casual observers. The researchers came up with several techniques for doing this, including subtle fading, camouflage graffiti, and camouflage art. Here's how the perturbed signs look when printed out as posters and stuck onto real signs:

Subtle perturbations cause a neural network to misclassify stop signs as speed limit 45 signs, and right turn signs as stop signs. IMAGES: EVTIMOV ET AL

And here are two attacks that are easier to manage on a real-world sign, since they're stickers rather than posters:

Camouflage graffiti and art stickers cause a neural network to misclassify stop signs as speed limit 45 signs or yield signs. IMAGES: EVTIMOV ET AL

Because the stickers have a much smaller area to work with than the posters, the perturbations they create have to be more significant, but it's certainly not obvious that they're not just some random graffiti. And they work almost as well. According to the researchers:

> *The Stop sign is misclassified into our target class of Speed Limit 45 in 100% of the images taken according to our evaluation methodology. For the Right Turn sign... Our attack reports a 100% success rate for misclassification with 66.67% of the images classified as a Stop sign and 33.7% of the images classified as an Added Lane sign. [The camouflage graffiti] attack succeeds in causing 73.33% of the images to be misclassified. In [the camouflage abstract art attack], we achieve a 100% misclassification rate into our target class.*

In order to develop these attacks, the researchers trained their own road sign classifier in TensorFlow using a publicly available, labeled

"white box" access to the classifier, meaning that they can't mess with its training or its guts, but that they can feed things in and see what comes out— like if you owned an autonomous car, and could show it whatever signs you wanted and see if it recognized them or not, a reasonable assumption to make. Even if you can't hack directly into the classifier itself, you could still use this feedback to create a reasonably accurate model of how it classifies things. Finally, the researchers take the image of the sign you want to attack and feed it plus their classifier into an attack algorithm that outputs the adversarial image for you. Mischief managed.

It's probably safe to assume that the classifiers used by autonomous cars will be somewhat more sophisticated and robust than the one that these researchers managed to fool so successfully. (It used only about 4,500 signs as training input.) It's probably **not** safe to assume that attacks like these won't *ever* work, though, because even the most sophisticated deep neural network-based algorithms can be really, really dumb at times for reasons that aren't always obvious. The best defense is probably for autonomous cars to use a multi-modal system for road sign detection, for the same reason that they use multi-modal systems for obstacle detection: It's dangerous to rely on just one sensor (whether it's radar, lidar, or cameras), so you use them all at once, and hope that they cover for each other's specific vulnerabilities. Got a visual classifier? Great, make sure and couple it with some GPS locations of signs. Or maybe add in something like a dedicated red octagon detection system. My advice, though, would just be to do away with signs all together, at the same time that you

completely to robots. Problem solved.

Robust Physical-World Attacks on Machine Learning Models, by Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song from the University of Washington, the University of Michigan Ann Arbor, Stony Brook University, and the University of California Berkeley, can be found on arXiv.

TAGS

AUTOMOTIVE SENSORS | SENSORS | NEURAL NETWORKS | CLASSIFICATION | AUTONOMOUS VEHICLES

ABOUT THE AUTHOR

Evan Ackerman is a senior editor at *IEEE Spectrum*. Since 2007, he has written over 6,000 articles on robotics and technology. He has a degree in Martian geology and is excellent at playing bagpipes.

READER RESPONSES SORT BY POPULAR

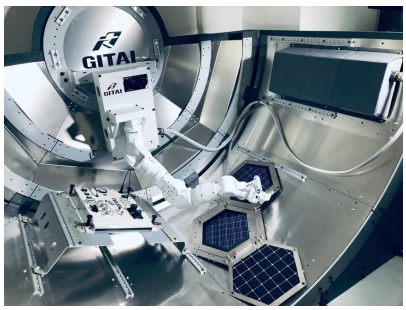
Add comment...

PUBLISH

READ ALSO

2021 Herz Award Goes to Former IEEE Senior Director of IT

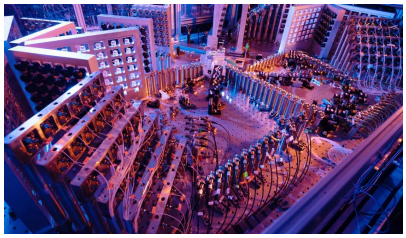
16 HOURS AGO | 2 MIN READ |



NEWS | ROBOTICS

GITAI's Autonomous Robot Arm Finds Success on ISS

19 HOURS AGO | 3 MIN READ |



NEWS | COMPUTING

Two of World's Biggest Quantum Computers Made in China

06 NOV 2021 | 2 MIN READ |

Related Stories

FEATURE | TRANSPORTATION

The Next Generation of AI-Enabled Cars Will Really Understand You

INTERVIEW | SENSORS

Event-Based Camera Chips Are Here, What's Next?

Autonomous Racing Drones Dodge Through Forests at 40 kph

FEATURE | TRANSPORTATION

2021 TOP 10 TECH CARS

The trend toward all-electric is accelerating

Slight Street Sign Modifications Can Completely Fool Machine Learning Algo...

BY LAWRENCE ULRICH

26 MAR 2021 | 1 MIN READ |

PHOTO: RIMAC AUTOMOBILI

The COVID-19 pandemic put the auto industry on its own lockdown in 2020. But the technological upheavals haven't slowed a bit.

powered, either in EV or gas-electric hybrid form. A few critical model introductions were delayed by the virus, including the debut of one of our boldface honorees: the long-awaited 2021 Lucid Air electric sedan. It's expected to hit the market in a few months. But the constellation of 2021's electric stars covers many categories and budgets, from the ultra-affordable, yet tech-stuffed Hyundai Elantra Hybrid to the US $2.4 million Rimac C Two hypercar.

Keep Reading ↓

NEWS | ROBOTICS

Video Friday: Your Robot Dog Your weekly selection of awesome robot videos

BY EVAN ACKERMAN

05 NOV 2021 | 3 MIN READ |

TAGS

VIDEO FRIDAY | ROBOTICS

Video Friday is your weekly selection of awesome robotics videos, collected by your friends at *IEEE Spectrum* robotics. We'll also be posting a weekly calendar of upcoming robotics events for the next few months; here's what we have so far (send us your events!):

ICRA 2022 - MAY 23-27, 2022 - PHILADELPHIA, PA, USA

Let us know if you have suggestions for next week, and enjoy today's videos.

Keep Reading ↓

Since its introduction in 1978, Master Bond EP29LPSP has been the epoxy compound of choice in a variety of challenging applications. Ideal for demanding cryogenic environments, two-part EP29LPSP can withstand temperatures as low as 4K and can resist cryogenic shock when, for instance, it is cooled from room temperature to cryogenic temperatures within a 5-10 minute window. Optically clear EP29LPSP has superior physical strength, electrical insulation, and chemical resistance properties. It also meets NASA low outgassing requirements and exhibits a low exotherm during cure. This low viscosity compound is easy to apply and bonds well to metals, glass, ceramics, and many different plastics. Curable at room temperature, EP29LPSP attains its best results when cured at 130-165°F for 6-8 hours.

In over a dozen published research articles, patents, and manufacturers' specifications, scientists and engineers have identified EP29LPSP for use in their applications due to its unparalleled performance in one or more areas. Table 1 highlights several commercial and research applications that use Master Bond EP29LPSP. Table 2 summarizes several patents that reference EP29LPSP. Following each table are brief descriptions of the role Master Bond EP29LPSP plays in each application or invention.

Keep Reading ↓

PROFILE | ARTIFICIAL INTELLIGENCE

This 6-Million-Dollar AI Changes Accents as You Speak

Three international Stanford undergrads start company to "help the world understand"

BY TEKLA S. PERRY

05 NOV 2021 | 4 MIN READ |

In 2020, Stanford students Shawn Zhang, Maxim Serebryakov, and Andres Perez Soderi [left to right] founded the AI-powered accent-translation company Sanas. SANAS

TAGS

ARTIFICIAL INTELLIGENCE | STARTUPS | SPEECH RECOGNITION TECHNOLOGY

Stanford University prides itself on its international diversity, touting that today's undergraduates hail from 70 countries. So a friend-group that included a computer science major from China, an AI-focused management science and engineering (MSE) major from Russia, and a business-oriented MSE major from Venezuela isn't an anomaly. The friends did the normal things Stanford students do with their free time, like fountain hopping, cheering at football games, and hiking the trail around the Stanford Dish radio telescope.

And then came the pandemic.

Keep Reading ↓

Fixing the Future

On *IEEE Spectrum*'s Fixing the Future podcast, host Steven Cherry talks with the brightest minds in technology about

challenges

All Fixing the Future episodes →

Solving the Electric Vehicle Charging Conundrum John Voelcker explains why getting EVs charged is easier than you think

EPISODE 33 | 36:24 | 26 OCT 2021 |

IBM's Fall From World Dominance Tech historian James Cortada has charted the company's many highs and lows—and thinks it's still a contender

EPISODE 32 | 25:55 | 11 AUG 2021 |

How Computers Can Finally Detect Sarcasm Ramya Akula and the tech that lets sentiment analysis spot mocking words

EPISODE 31 | 19:29 | 01 JUL 2021 |

Swapping Electricity for Chemicals Sunthetics' CEO Myriam Sbeiti attack on industrial chemistry's sustainability problem

EPISODE 30 | 18:30 | 21 JUN 2021 |

Seaborg Technologies CEO, Troels Schönefeldt, on why this could be a good idea

EPISODE 29 | 24:04 | 14 JUN 2021 |

Can Data Make Fossil Fuels More Efficient? Data-science startup Cognite's Carolina Torres discusses how to help energy firms reduce production losses and improve sustainability

EPISODE 28 | 19:44 | 03 JUN 2021 |

FEATURE | TRANSPORTATION

2021'S TOP TEN TECH CARS: PORSCHE 911 TURBO S

Fast as lightning, stable when wet

Slight Street Sign Modifications Can Completely Fool Machine Learning Algo...

BY LAWRENCE ULRICH

26 MAR 2021 | 2 MIN READ |

PHOTO: PORSCHE

Among modern sports cars, there's fast, and then there's the Porsche 911 Turbo. I've driven many generations of this Autobahn brawler, but I still wasn't quite prepared for what Porsche could do with 477 kilowatts—640 horsepower.

Keep Reading ↓

OPINION | BIOMEDICAL

We've Entered a New Era of Streaming Health Care. Now What?

COVID-19 forced the transition to digital medicine, but there's much still to do

BY LESLIE SAXON DEVIN SKOLL

05 NOV 2021 | 6 MIN READ |

ISTOCKPHOTO

TAGS

HEALTH TRACKERS | HEALTH CARE | TELECOMMUNICATIONS | ARTIFICIAL INTELLIGENCE | TELEMEDICINE

It's a rare soul who truly enjoys going to the doctor. Many people feel vulnerable when seated on a chilly exam table in that paper robe, and fear of the unknown or denial can keep people from seeking vital care. This problem has grown in importance as research has shown that preventive care and early detection of disease improve outcomes.

The COVID-19 pandemic changed the equation. It forced a transition to virtual health care delivery and exposed many patients, providers, and health care delivery organizations to the efficiencies and qualities of telemedicine for the first time. The pandemic also pushed venture capital investments in digital health to an all-time high; in 2021, investments had reached $14.7 billion by the end of the second quarter, surpassing the figure for all of 2020.

Keep Reading ↓

Simulation Apps At Work: 4 Use Cases Specialized simulation apps enable collaboration across the enterprise and drive innovation

BY COMSOL

23 SEP 2021 | 1 MIN READ |

TAGS

SIMULATION | COMSOL

Organizations are turning to specialized simulation apps to enable collaboration between engineers across the enterprise. This white paper covers the underlying technology for creating and deploying simulation apps to larger groups of people. Use cases highlight how apps are being used to benefit product development and drive innovation.

NEWS | THE INSTITUTE

Novel Approaches for Forecasting Electricity Demand Researchers offer ways to make more accurate predictions post-COVID

BY KATHY PRETZ

04 NOV 2021 | 4 MIN READ |

ISTOCKPHOTO

TAGS

IEEE NEWS | IEEE DATAPORT | ELECTRICITY | UTILITIES | COVID-19 | CORONAVIRUS | DATA SETS | GRID | IEEE PRODUCTS SERVICES

Among the many industries impacted by the COVID-19 pandemic were electric utilities. Demand for electricity dropped last year in nearly all countries, according to the International Energy Association.

The closing of office buildings, schools, factories, and other facilities made it challenging for utilities to forecast how much electricity customers would be consuming. Utilities base some of their predictions on historical data such as weather and atmospheric conditions, holidays, economic events, and geographic information. But no comparative data existed for the lockdowns that took place around the world.

complete shutdowns are still occurring. Many employees continue to work from home. The fluid situation has left utility companies scrambling for solutions to improve load-forecasting accuracy.

Keep Reading ↓