Protecting computer vision from adversarial attacks

June 17 2022, by Holly Ober

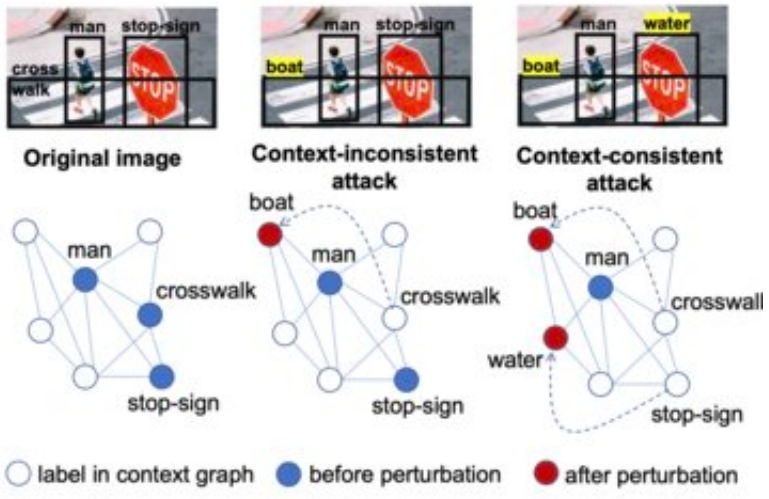

Illustration showing how an attacker could cause a computer vision system to miscategorize objects it sees through the camera. Mislabeling one object might not be enough to make a bad decision but mislabeling several related objects will. Credit: Cai et.al. 2022

Advances in computer vision and machine learning have made it possible for a wide range of technologies to perform sophisticated tasks with little or no human supervision. From autonomous drones and self-driving cars to medical imaging and product manufacturing, many computer applications and robots use visual information to make critical decisions. Cities increasingly rely on these automated technologies for public safety and infrastructure maintenance.

However, compared to humans, computers see with a kind of tunnel vision that leaves them vulnerable to attacks with potentially catastrophic results. For example, a human driver, seeing graffiti covering a stop sign, will still recognize it and stop the car at an intersection. The graffiti might cause a self-driving car, on the other hand, to miss the stop sign and plow through the intersection. And, while human minds can filter out all sorts of unusual or extraneous visual information when making a decision, computers get hung up on tiny deviations from expected data.

This is because the brain is infinitely complex and can process multitudes of data and past experiences simultaneously to arrive at nearly instantaneous decisions appropriate for the situation. Computers rely on mathematical algorithms trained on datasets. Their creativity and cognition are constrained by the limits of technology, math, and human foresight.

Malicious actors can exploit this vulnerability by changing how a computer sees an object, either by altering the object itself or some aspect of the software involved in the vision technology. Other attacks can manipulate the decisions the computer makes about what it sees. Either approach could spell calamity for individuals, cities, or companies.

A team of researchers at UC Riverside's Bourns College of Engineering are working on ways to foil attacks on computer vision systems. To do that, Salman Asif, Srikanth Krishnamurthy, Amit Roy-Chowdhury, and Chengyu Song are first figuring out which attacks work.

"People would want to do these attacks because there are lots of places where machines are interpreting data to make decisions," said Roy-Chowdhury, the principal investigator on a recently concluded DARPA AI Explorations program called Techniques for Machine Vision Disruption. "It might be in the interest of an adversary to manipulate the

data on which the machine is making a decision. How does an adversary attack a data stream so the decisions are wrong?"

An adversary would inject some malware into the software on a self-driving vehicle, for example, so that when data comes in from the camera it is slightly perturbed. As a result, the models installed to recognize a pedestrian fail and the system would be hallucinating an object or not seeing one that does exist. Understanding how to generate effective attacks helps researchers design better defense mechanisms.

"We are looking at how to perturb an image so that if it is analyzed by a machine learning system, it is miscategorized," Roy-Chowdhury said. "There are two main ways to do this: Deepfakes where the face or facial expressions of someone in a video have been altered so as to fool a human, and adversarial attacks in which an attacker manipulates how the machine makes a decision but a human is usually not mistaken. The idea is you make a very small change in an image that a human can't perceive but that an automated system will and make a mistake."

Roy-Chowdhury, his collaborators, and their students have found that the majority of existing attack mechanisms are targeted toward misclassifying specific objects and activities. However, most scenes contain multiple objects and there is usually some relationship among the objects in the scene, meaning certain objects co-occur more frequently than others.

People who study computer vision call this co-occurrence "context." Members of the group have shown how to design context-aware attacks that alter the relationships between objects in the scene.

"For example, a table and chair are often seen together. But a tiger and chair are rarely seen together. We want to manipulate all of these together," said Roy-Chowdhury. "You could change the stop sign to a

speed limit sign and remove the crosswalk. If you replaced the stop sign with a speed limit sign but left the crosswalk, the computer in a self-driving car might still recognize it as a situation where it needs to stop."

Earlier this year, at the Association for the Advancement of Artificial Intelligence conference, the researchers showed that in order for a machine to make a wrong decision it's not enough to manipulate only one object. The group developed a strategy to craft adversarial attacks that change multiple objects simultaneously in a consistent manner.

"Our main insight was that successful transfer attacks require holistic scene manipulation. We learn a context graph to guide our algorithm on which objects should be targeted to fool the victim model, while maintaining the overall scene context," said Salman Asif.

In a paper presented this week at the Conference on Computer Vision and Pattern Recognition conference, the researchers, along with their collaborators at PARC, a research division of the Xerox company, build further on this concept and propose a method where the attacker had no access to the victim's computer system. This is important because with each intrusion the attacker risks detection by the victim and a defense against the attack. The most successful attacks are therefore likely to be ones that do not probe the victim's system at all, and it is crucial to anticipate and design defenses against these "zero-query" attacks.

Last year, the same group of researchers exploited contextual relationships in time to craft attacks against video sequences. They used geometric transformations to design very efficient attacks on video classification systems. The algorithm leads to successful perturbations in surprisingly few attempts. For example, adversarial examples generated from this technique have better attack success rates with 73% fewer attempts compared to state-of-the-art methods for video adversarial attacks. This allows for faster attacks with far fewer probes into the

victim system. This paper was presented at the premier machine learning conference, Neural Information Processing Systems 2021.

The fact that context-aware adversarial attacks are much more potent on natural images with multiple objects than existing ones that mostly focus on images with a single dominant object opens the route to more effective defenses. These defenses can consider the contextual relationships between objects in an image, or even between objects across a scene in images by multiple cameras. This holds the potential for the development of significantly more secure systems in the future.

More information: Zikui Cai et al, Context-Aware Transfer Attacks for Object Detection. arXiv:2112.03223v1 [cs.CV], arxiv.org/pdf/2112.03223.pdf

Zikui Cai et al, Zero-Query Transfer Attacks on Context-Aware Object Detectors. arXiv:2203.15230v1 [cs.CV], arxiv.org/pdf/2203.15230.pdf

Shasha Li et al, Adversarial Attacks on Black Box Video Classifiers: Leveraging the Power of Geometric Transformations. arXiv:2110.01823v2 [cs.CV], arxiv.org/pdf/2110.01823.pdf

Provided by University of California - Riverside