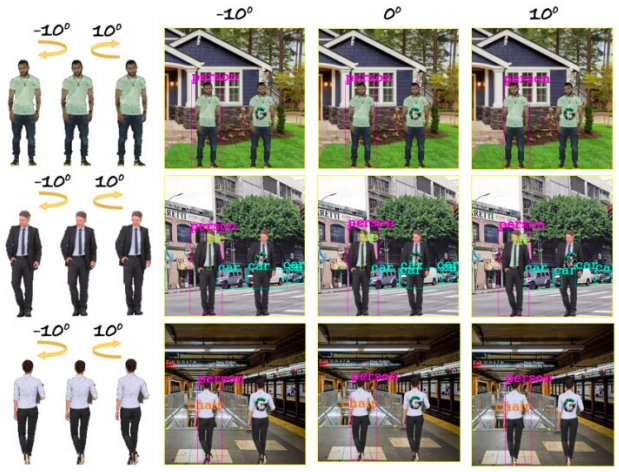
Fooling deep neural networks for object detection with adversarial 3-D logos

30 July 2020, by Ingrid Fadelli



Examples of the researchers' 3D adversarial logo attack using different 3D object meshes, with the aim of fooling a YOLOV2 detector. Credit: Chen et al.

Over the past decade, researchers have developed a growing number of deep neural networks that can be trained to complete a variety of tasks, including recognizing people or objects in images. While many of these computational techniques have achieved remarkable results, they can sometimes be fooled into misclassifying data.

An adversarial attack is a type of cyberattack that specifically targets deep neural networks, tricking them into misclassifying data. It does this by creating adversarial data that closely resembles and yet differs from the data typically analyzed by a deep neural network, prompting the network to make incorrect predictions, failing to recognize the slight differences between real and adversarial data.

In recent years, this type of attack has become increasingly common, highlighting the vulnerabilities and flaws of many deep neural networks. A specific type of adversarial attack that has emerged in recent years entails the addition of adversarial patches (e.g., logos) to images. This attack has so far primarily targeted models that are trained to detect objects or people in 2-D images.

Researchers at Texas A&M University, University of Texas at Austin, University of Science and Technology in China, and the MIT-IBM Watson AI Lab have recently introduced a new attack that entails the addition of 3-D adversarial logos to images with the aim of tricking deep neural networks for object detection. This attack, presented in a paper pre-published on arXiv, could be more applicable to real-world situations, as most real data processed by deep neural networks is in 3-D.

"The primary aim of this work is to generate a structured patch in an arbitrary shape (called a 'logo' by us), termed as a 3-D adversarial logo that, when appended to a 3-D human mesh and then rendered into 2-D images, can consistently fool the object detector under different human postures," the researchers wrote in their paper.

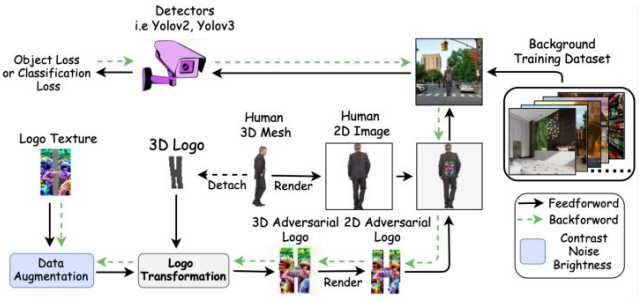


Figure outlining how the attack presented by the researchers works. Credit: Chen et al

Essentially, the researchers created an arbitrary

shape logo based on a pre-existing 2-D texture image. Subsequently, they mapped this image onto a 3-D adversarial logo, employing a texture-mapping method known as logo transformation. The 3-D adversarial logo they crafted could then serve as an adversarial texture, allowing the attacker to easily manipulate its shape and position.

In contrast with previously introduced attacks that utilize adversarial patches, this new type of attack maps logos in 3-D, yet it derives its shapes from 2-D images. As a result, it enables the creation of versatile adversarial logos that are can trick a broad variety of object or person detection methods, including those used in real-world situations, such as techniques for identifying people in CCTV footage.

"We render 3-D meshes with the 3-D adversarial logo attached into 2-D scenarios and synthesize images that could fool the detector," the researchers wrote in their paper. "The shape of our 3-D adversarial logo comes from the selected logo texture in the 2-D domain. Hence, we can perform versatile adversarial training with shape and position controlled."

The researchers tested the success rate of their adversarial logo attack by implementing it on two state-of-the-art deep neural network-based object detectors, known as YOLOv2 and YOLOv3. In these evaluations, the 3-D adversarial logo fooled both detectors robustly, causing them to misclassify images taken from a variety of angles and in which humans were in different postures.

These results confirm the vulnerabilities of deep neural network-based techniques for detecting objects or humans in images. They thus further highlight the need to develop deep learning methods that are better at spotting adversarial images or logos and that are harder to fool using synthesized data.

More information: Chen et al., Can 3-D adversarial logos cloak humans? arXiv:2006.14655 [cs.LG]. arxiv.org/abs/2006.14655