Adversarial attacks on surveillance cameras with 2D prints

As the field of artificial intelligence advances, surveillance cameras has been given digital brains to match their eyes. Machine-learning algorithms let them analyze live video with no humans necessary. For example, a smart camera called “AI Guardman” [1] is designed to help shop owners in Japan spot potential shoplifters. The system tries to match this pose data to predefined ‘suspicious’ behavior. If it sees something noteworthy, it alerts shopkeepers via a connected app. The company that produces this intelligent camera reports that in stores where they introduced their product, shoplifting damage is reduced from 3.5 million yen per year to 2 million yen [1]. As a result, this could be good news for public safety, helping police and first responders more easily spot crimes and accidents.

Although the application of artificial intelligence in automated surveillance cameras is promising, researchers are also trying to discover flaws in the detection system. Academics from a Belgian university have devised a method that uses a simple 2D image that can be printed on shirts or bags to make wearers invisible to camera surveillance systems that rely on machine learning to recognize humans in live video feeds [2]. In order to be invisible from the surveillance camera, the 2D patch must be placed around the middle of a person's "detection box" and must face the surveillance camera at all times (Figure 1 [2]).

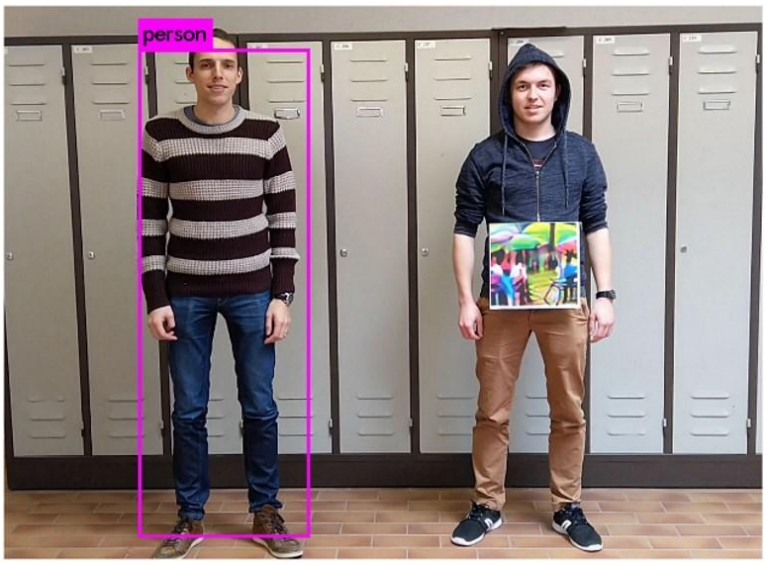

Figure 1: 2D image that is able to hide persons from a person detector

The method of providing an ML model with a spurious input that fools it into producing a wrong result is called adversarial attack [4]. Recently, Deep Neural Networks (DNN) have been found vulnerable to well-designed input samples. Because Deep Neural Networks are highly expressive models, it causes them to be unstable, which means that a small perturbation on the input could change the network's prediction. Taking advantage of this property, it is possible for the attacker to modify the input and obtain the target class as the output. There are two forms of adversarial attack: Adversarial noise and Adversarial patch [4].

On the one hand, by adding an adversarial noise, it means adding some pixels to the original picture to trick the neural network into predicting another class. We can carefully manipulate the noise so that the amplitude of the noise is minimized. The noise is usually extremely subtle to us so that we do not realize that the picture has been maliciously modified. One example is shown in Figure 2.

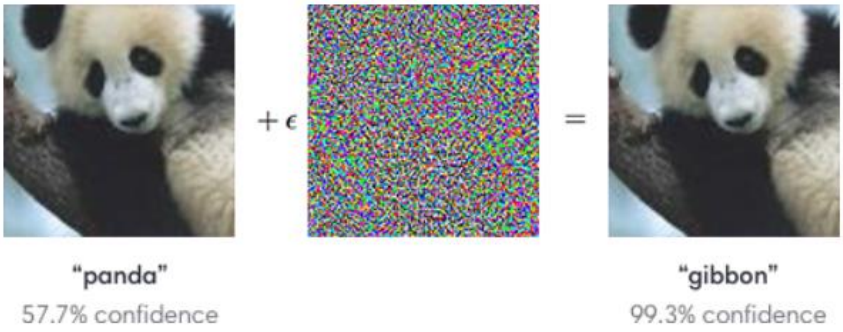

Figure 2: An adversarial input, overlaid on a typical image,

causes a classifier to miscategorize a panda as a gibbon.

On the other hand, adversarial patch is superimposed into the original image during attack. When the patched image is fed into neural network, the patch forces it to ignore other parts of the image and predict the target class as the attacker wants. It allows attackers to create a physical-world attack without prior knowledge of the lighting conditions, camera angle, or even the other items within the scene. An explanation for this is that an image may contain several items, but a classifier outputting only one target output has to determine which is the most 'salient' item in the frame [4]. The adversarial patch exploit this by getting all the attention of the classifier. One example is Figure 3 [4].

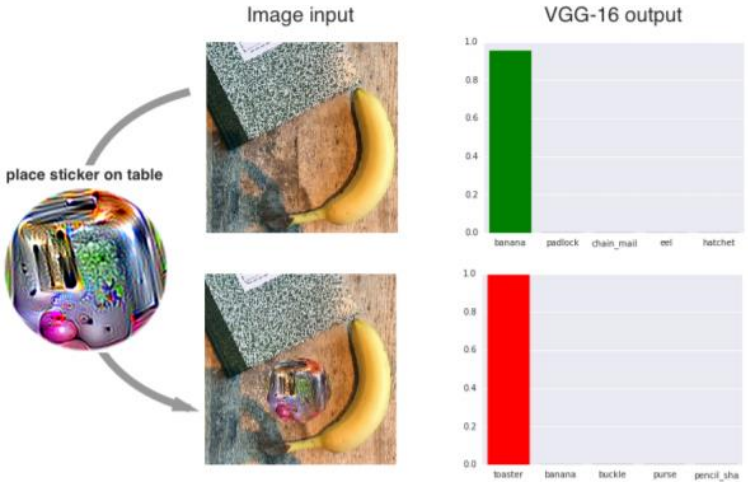

Figure 3: If we physically place a sticker targeted to the class "toaster" on the table,

the photograph is classified as a toaster with 99% confidence

The camera target in their research is called YOLOv2, an object detector using Convolutional Neural Network. The detector divides the image into regions and makes three predictions: bounding boxes (an area that is distinct from the surrounding), object score (how likely the region contains an object) and class score (the probability for each of the classes available) for each region. These bounding boxes are weighted by the object scores. By setting the probability threshold, the boxes high probability are highlighted and this produces the final output. Figure 4 [3] is an illustration.

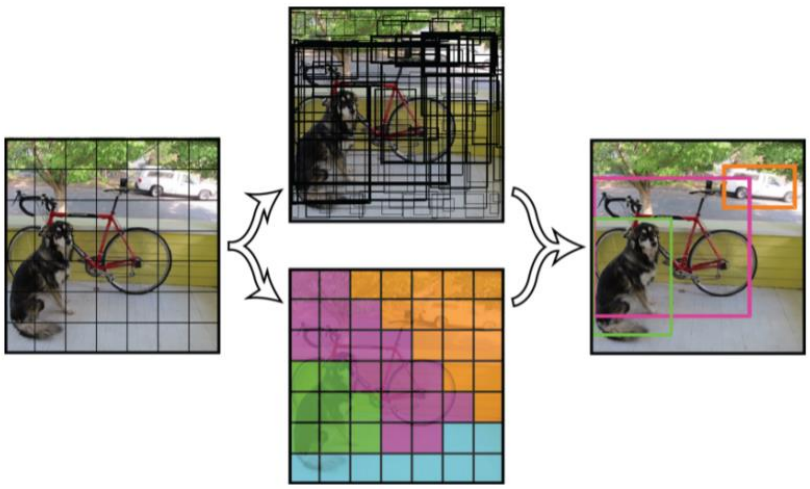

Figure 4: overview of YOLOv2 architecture

In order to produce the patch that hides person from the detector, they experimented three approaches [3]: minimize the score of the class person (the probability that the classifier considers a part of the image as a person); minimize the object score (the probability that the classifier considers a part of image as a general object); minimize a combination of both.

For each of the three approach, they generate the 2D image patch by using an optimization process [3]: First, applying patch to real images of different people in the dataset. After feeding the image into the detector and get the output, measure the score of

persons that are still detected, which are used to calculate a loss function. Then using the loss function, the optimizer modifies the patch further in order to fool the detector even more.

The graph below (Figure 5 [3]) shows the effectiveness of the three approaches: minimizing score of class person (purple line CLS), object score (red line OBJ), or both (green line OBJ-CLS). AP denotes the resulting average precision. The conclusion is the red curve, which represents minimizing the object score only, reduces the accuracy of the detector most effectively.

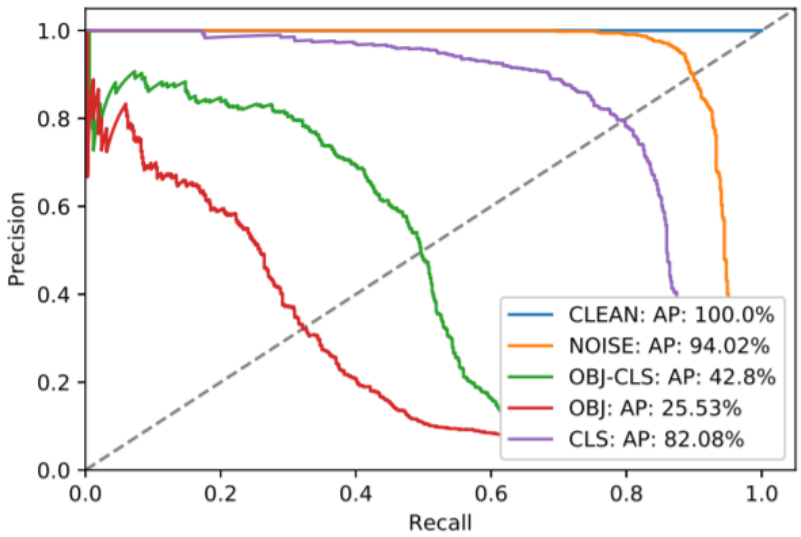

Figure 5: Effectiveness of three approaches,

Minimizing only object score has the best performance

This research raises questions regarding the reliability of automated security systems. The classification model that they rely on, which is neural network, could be exploited for malicious purposes. For instance, an attack that could be used to circumvent surveillance systems, intruders can sneak around undetected, in theory, by wearing a T-shirt containing this kind of adversarial patches in front of their body aimed towards the

surveillance camera. The same system can also be adapted to hide certain objects from view. For example, a "patch" could hide cars or bags from view as well, if the surveillance system is configured to detect certain objects instead of humans.

Apart from being a threat to automated security systems, the result of this kind of attack encourages us to reevaluate other systems as well. One example is the Tesla Autopilot [5], some researchers has successfully modified the stop signs to fool the classifier. Figure 6 [5] shows a modified stop sign which is classified as a sports ball. Also, they have managed to make the classifier considering the stop sign as a bird, a person, a clock, etc.

Figure 6: modified stop sign classified as sports ball by Tesla Autopilot

Works cited:

[1] J. Vincent, “This Japanese AI security camera shows the future of surveillance will be automated,” The Verge, 26-Jun-2018. [Online]. Available: https://www.theverge.com/2018/6/26/17479068/ai-guardman-security-camera-shoplifter-japan-automated-surveillance . [Accessed: 20-May-2019].

[2] C. Cimpanu, “Academics hide humans from surveillance cameras with 2D prints,” ZDNet, 05-May-2019. [Online]. Available: https://www.zdnet.com/article/academics-hide-humans-from-surveillance-cameras-with-2d-prints/ . [Accessed: 20-May-2019].

[3] S. Thys, W. V. Ranst, and T. Goedeme, “Fooling automated surveillance cameras: adversarial patches to attack person detection,” arXiv:1904.08653v1, 2019.

[4] S. Joglekar, “Adversarial patches for CNNs explained,” codeburst, 09-Jan-2018. [Online]. Available: https://codeburst.io/adversarial-patches-for-cnns-explained-d2838e58293 . [Accessed: 20-May-2019].

[5] S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau, “ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector, ” arXiv:1804.05810v3, 2019