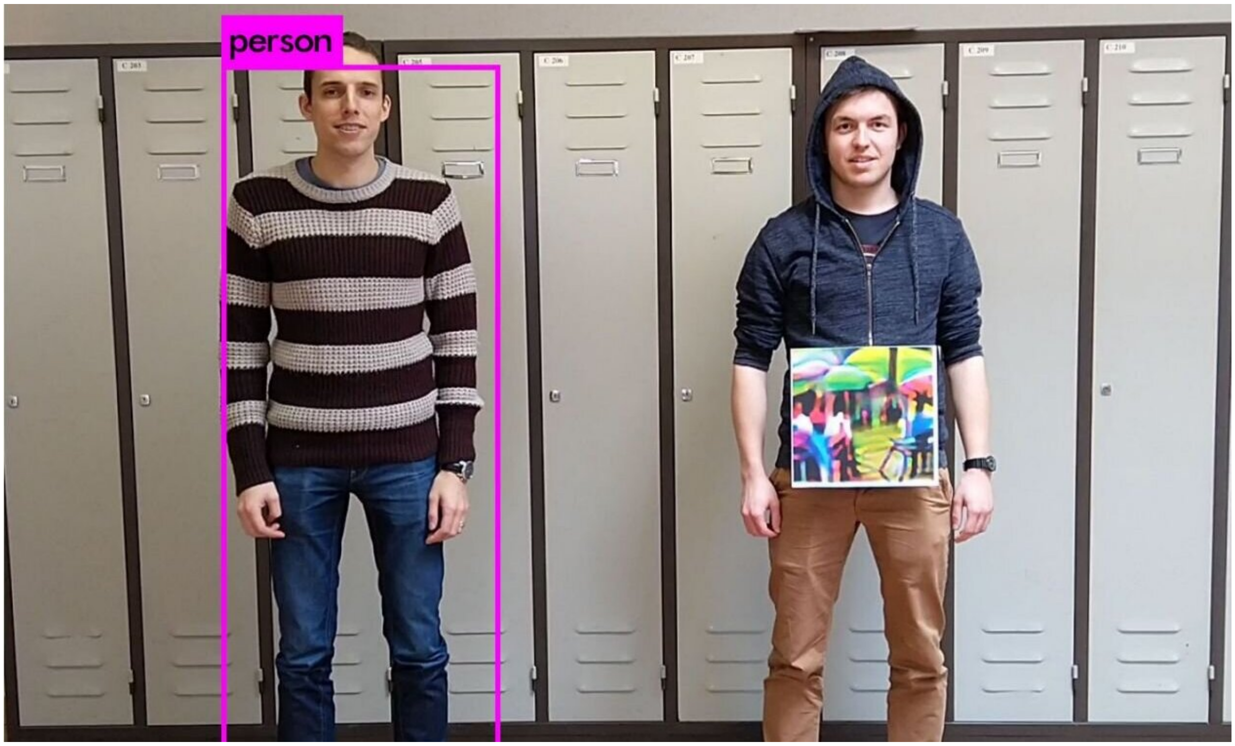
Using a printed adversarial patch to fool an AI system

April 24 2019, by Bob Yirka



Left: The person without a patch is successfully detected. Right: The person holding the patch is ignored. Credit: arXiv:1904.08653 [cs.CV]

A trio of researchers at the University of KU Leuven in Belgium has found that it is possible to confuse an AI system by printing a certain picture and holding it against their body as the AI system tries to identify them as a human being. Simen Thys, Wiebe Van Ranst and Toon

Goedemé have written a paper describing their efforts and have uploaded it to the *arXiv* preprint server. They have also posted a video on YouTube showing what they accomplished.

For an AI system to learn something, such as identifying objects (including human beings) in a scene, it must be trained—the training involves showing it thousands of objects that fit into given categories until general patterns emerge. But as prior research has suggested, such systems can sometimes become confused if they are presented with something they were not trained to see. In this case, a 2D picture of people holding colorful umbrellas. Such AI fooling images are known as adversarial patches.

As AI systems become more accurate and sophisticated, governments and corporations have started using them for real-world applications. One well-known application used by governments is spotting individuals that might stir up trouble. Such systems are trained to recognize the human form—once that happens, a facial recognition system can be activated. Recent research has shown that facial recognition systems can be fooled by users wearing specially designed eyeglasses. And now it appears that human-spotting AI systems can be fooled by images placed in front of their forms.

In their attempt to fool one particular human-recognizing AI system called YoLo(v2) the researchers created or edited various types of images which they then tested with the AI system until they found one that worked particularly well—an image of people holding colorful umbrellas that had been altered by rotating it and adding noise. To fool the AI system, the photograph was held in a position that occupied the box that the AI system constructed to determine if a given object was identifiable.

The researchers demonstrated the effectiveness of their adversarial patch by creating a video that showed the boxes drawn by the AI system as it encountered objects in its field of view and then posted identifying labels to them. Without the patch, the system very easily identified people in the video as human beings—but if one of them held the patch over their mid-section, the AI system was no longer able to detect their presence.

More information: Simen Thys et al. Fooling automated surveillance cameras: adversarial patches to attack person detection, arXiv:1904.08653 [cs.CV] arxiv.org/abs/1904.08653

Citation: Using a printed adversarial patch to fool an AI system (2019, April 24) retrieved 30 April 2024 from https://techxplore.com/news/2019-04-adversarial-patch-ai.html