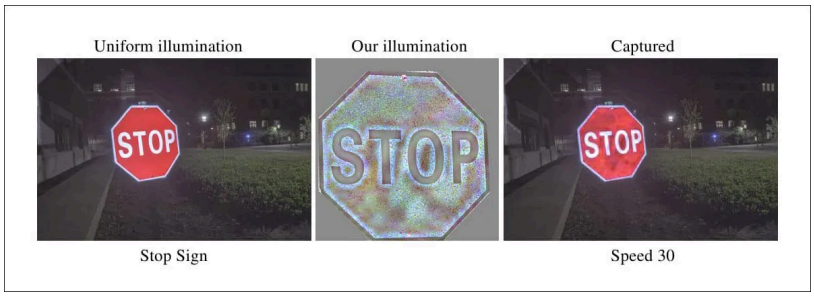
Optical Adversarial Attack Can Change the Meaning of Road Signs

Originally published at https://www.unite.ai/optical-adversarial-attack-can-change-the-meaning-of-road-signs/

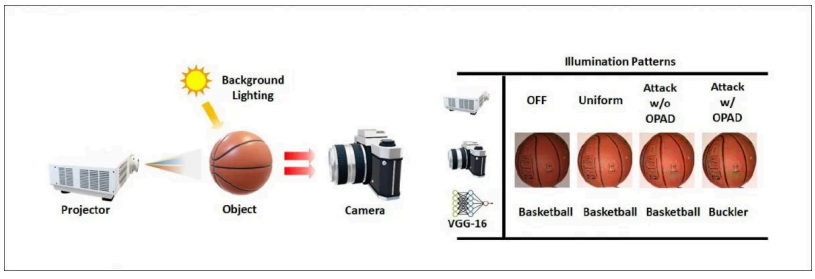
Researchers in the US have developed an adversarial attack against the ability of machine learning systems to correctly interpret what they see – including mission-critical items such as road signs – by shining patterned light onto real world objects. In one experiment, the approach succeeded in causing the meaning of a 'STOP' roadside sign to be transformed into a '30mph' speed limit sign.



Perturbations on a sign, created by shining crafted light on it, distorts how it is interpreted in a machine learning system. Source: https://arxiv.org/pdf/2108.06247.p

The research is entitled *Optical Adversarial Attack*, and comes from Purdue University in Indiana.

An OPtical ADversarial attack (OPAD), as proposed by the paper, uses structured illumination to alter the appearance of target objects, and requires only a commodity projector, a camera and a computer. The researchers were able to successfully undertake both white-box and black box attacks using this technique.

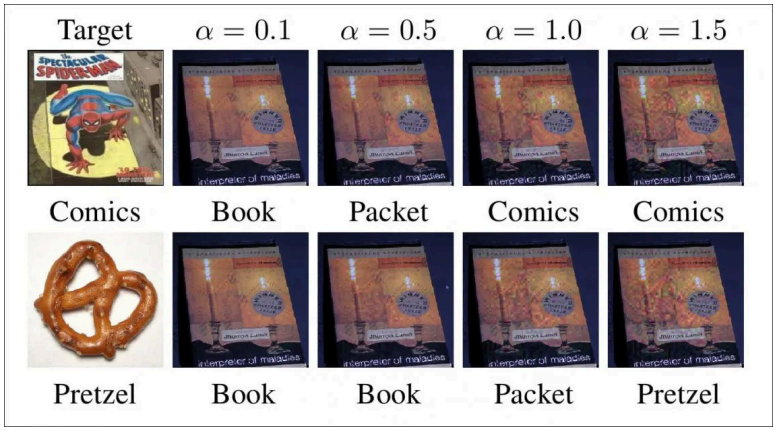


The OPAD set-up, and the minimally-perceived (by people) distortions that are adequate to cause a misclassification.

The set-up for OPAD consists of a ViewSonic 3600 Lumens SVGA projector, a Canon T6i camera, and a laptop computer.

Black Box and Targeted Attacks

White box attacks are unlikely scenarios where an attacker may have direct access to a training model procedure or to the governance of the input data. Black box attacks, conversely, are typically formulated by inferring how a machine learning is composed, or at least how it behaves, crafting 'shadow' models, and developing adversarial attacks designed to work on the original model.



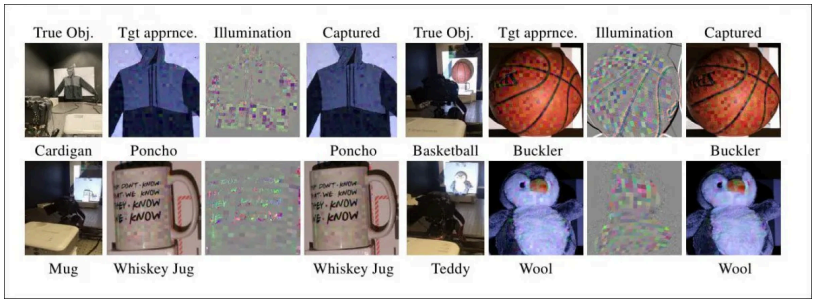
Here we see the amount of visual perturbation necessary to fool the classifier.

In the latter case, no special access is needed, though such attacks are greatly aided by the ubiquity of open source computer vision libraries and databases in current academic and commercial research.

All the OPAD attacks outlined in the new paper are 'targeted' attacks, which specifically seek to alter how certain objects are interpreted. Though the system has also been demonstrated capable of achieving generalized, abstract attacks, the researchers contend that a real-world attacker would have a more specific disruptive objective.

The OPAD attack is simply a real-world version of the frequently-researched principle of injecting noise into images that will be used in computer vision systems. The value of the approach is that one can simply 'project' the perturbations onto the target object in order to trigger the misclassification, whereas ensuring that 'Trojan horse' images end up in the training process is rather harder to accomplish.

In the case where OPAD was able to impose the hashed meaning of the 'speed 30' image in a dataset onto a 'STOP' sign, the baseline image was obtained by lighting the object uniformly at a 140/255 intensity. Then projector-compensated illumination was applied as a projected gradient descent attack.



Examples of OPAD misclassification attacks.

The researchers observe that the main challenge of the project has been to calibrate and set up the projector mechanism so that it achieves a clean 'deception', since angles, optics and several other factors are a challenge to the exploit.

Additionally, the approach is only likely to work at night. Whether the obvious illumination would reveal the 'hack' is also a factor; if an object such as a sign is already illuminated, the projector must compensate for that illumination, and the amount of reflected perturbation also needs to be resistant to headlights. It would seem to be a system that would work best in urban environments, where environmental lighting is likely to be more stable.

The research effectively builds an ML-oriented iteration of Columbia University's 2004 research into changing the appearance of objects by project other images onto them – an optics-based experiment that lacks the malign potential of OPAD.

In testing, OPAD was able to fool a classifier for 31 out of 64 attacks – a 48% success rate. The researchers note that the success rate depends greatly on the type of object being attacked. Mottled or curved surfaces (such as, respectively, a teddy bear and a mug) cannot provide enough direct reflectivity to perform the attack. On the other hand, intentionally reflective flat surfaces such as road signs are ideal environments for an OPAD distortion.

Open Source Attack Surfaces

All the attacks were undertaken against a specific set of databases: the German Traffic Sign Recognition Database (GTSRB, called GTSRB-CNN in the new paper), which was used to train the model for a similar attack scenario in 2018; the ImageNet VGG16 dataset; and the ImageNet Resnet-50 set.

So, are these attacks 'merely theoretical', since they are aimed at open source datasets, and not at the proprietary, closed systems in autonomous vehicles? They would be, if the major research arms did not rely on the open source ecostructure, including algorithms and datasets, and instead toiled in secret to produce closed-source datasets and opaque recognition algorithms.

But in general, that isn't how it works. Landmark datasets become the benchmarks against which all progress (and esteem/acclaim) becomes measured, while open source image recognition systems such as the YOLO series streak ahead, through common global cooperation, of any internally developed, closed system intended to operate on similar principles.

The FOSS Exposure

Even where the data in a computer vision framework will eventually be substituted with entirely closed data, the weights of the 'emptied out' models are still frequently calibrated in the early stages of development by FOSS data that will never be entirely discarded – which means that the resulting systems can potentially be targeted by FOSS methods.

Additionally, relying on an open source approach to CV systems of this nature makes it possible for private companies to avail themselves, free, of branched innovations from other global research projects, adding a financial incentive to keep the architecture accessible. Thereafter they can attempt to close the system only at the point of commercialization, by which time an entire array of inferable FOSS metrics are deeply embedded in it.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More