Impact of Fooling the Eyes of Autonomous Vehicles

Ahmed Fouad 23' (Electrical Engineering), Dr. Mooi Choo Chuah

(Computer Science Department)

Background

- YOLO (You Only Look Once) is an algorithm used for object recognition in real time. It is used widely for applications that require the detection of objects like traffic signs, people, and animals.
- Yolo algorithm uses convolutional neural networks (CNNs) to detect different object classes by a single forward propagation through the neural network. Then it predicts the classes probabilities and then draws the bounding boxes.
- Adversarial attacks aim to change the class prediction of the object detection algorithm by altering some of the pixel values in the image.
- By knowing the neural network model used in YOLO, adversarial perturbations ca be launched and trick the object detection model to identify object with wrong class.
- The goal is to generate adversarial images that confuses the neural network without being visible for human eye.

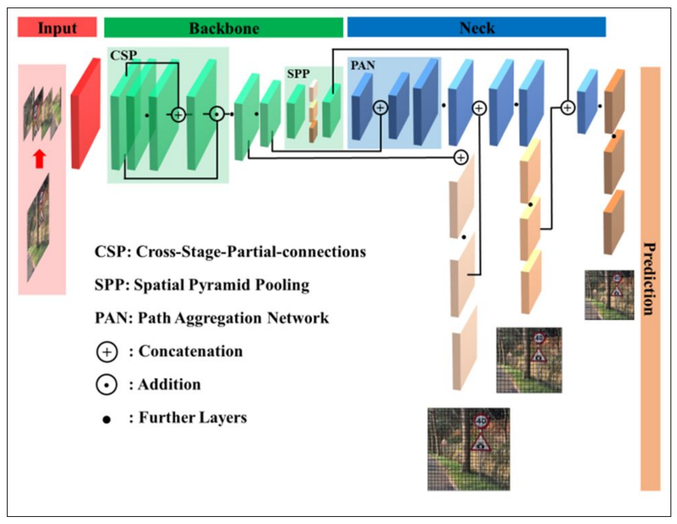

Figure 1. the architecture of YOLO[2].

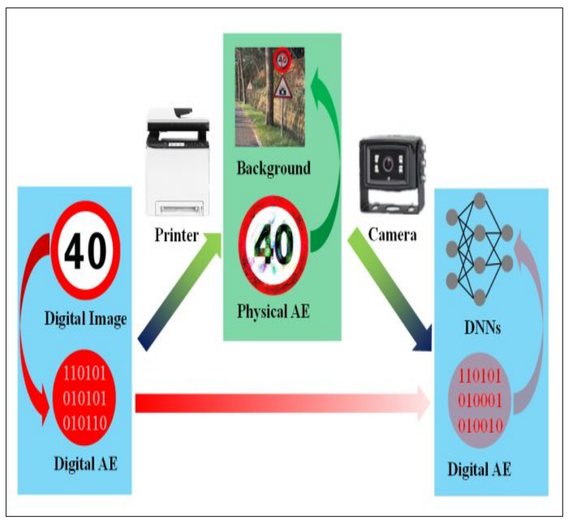

Figure 2. The Digital-Physical-Digital conversion of the adversarial images.

Method

- An F1Tenth autonomous car with an Intel Realsense camera was used in conjunction with Darknet_ros[1] software package to do the object detection.
- The configuration and weights files were based on YOLOv2. The attacks were performed on stop signs with varying perturbation levels and the car was programmed to stop if it detects the stop sign and becomes within sufficient distance from it.
- An Appearance Attack is done which makes the object detector misrecognize the attacked stop sign.
- We tested with attacked images with varying amount of noise to assess the vulnerability of the ROS version of YOLOv2.

Figure 3 Normal stop sign.

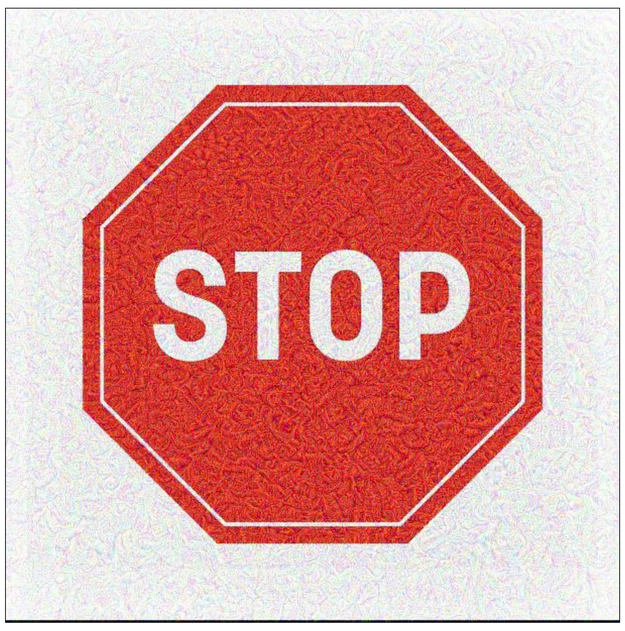

Figure 4 Stop sign after applying perturbations..

Results

- The car was able to recognize the normal stop sign and stop within 1.2 meters.
- When the attacked image is at a distance more than 1.2 meters, it either gets misrecognized to another class (a pizza, toaster), or does not get recognized at all.
- Factors such as illumination, distance from the attacked image, and angle of the image affect whether the attacked image gets recognized or not[2].

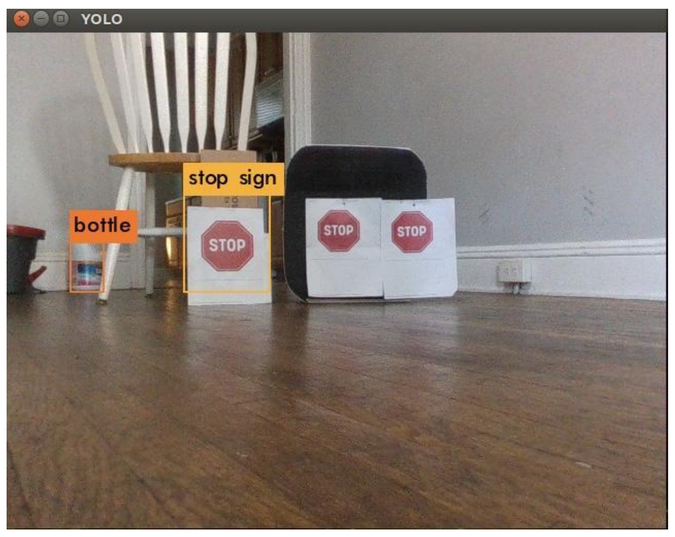

Figure 5 Attacked images not recognized by YOLOv2.

Conclusion

- Attacking the object detection system of autonomous vehicles proved to be dangerous. It can lead the car to make wrong decisions which jeopardize the safety of passengers.
- Object detection models used in autonomous vehicles must learn to adapt to such attacks and identify them in order to reach higher levels of autonomy.

Works Cited

- M. Bjelonic, YOLO ROS: Real-Time Object Detection for ROS, 2018, https://github.com/leggedrobotics/darknet_ros.
- Jia, Wei, et al. Fooling the Eyes of Autonomous Vehicles: Robust Physical Adversarial Examples Against Traffic Sign Recognition Systems. arXiv:2201.06192, arXiv, 16 Jan. 2022. arXiv.org, http://arxiv.org/abs/2201.06192.

Acknowledgments

We would like to thank Dr. Corey Montella and Zhihao Zheng for their technical help.
This poster was funded by STEM-SI program.

LEHIGH UNIVERSITY | LEHIGH UNIVERSITY PO

F1 TENTH f1tenth.org