Detecting Physical Adversarial Patch Attacks with Object Detectors

Damon Washington, E. Rebecca Caldwell

Dwashington124@rams.wssu.edu, caldwellr@wssu.edu

Winston-Salem State University

Winston-Salem, North Carolina

Abstract

Deep learning-based object detection technologies, such as YOLOv5 and Faster R-CNN, are being increasingly applied in safety-critical areas, including self-driving cars, surveillance systems, and smart transportation infrastructure. Although these models show remarkable performance under standard conditions, they are vulnerable to physical adversarial patch attacks. These attacks involve the careful placement of specifically designed printed perturbations in a scene to provoke misclassification or to obscure objects. In contrast to digital attacks that take place in controlled settings, physical adversarial patches operate under real-world conditions, where variables such as changing light, distance, angle, and occlusion can greatly influence their effectiveness, making them a significant danger.

This study explores detection-based defense mechanisms designed to identify physical adversarial patch attacks using object detectors. We assess various methods, including confidence-score analysis, monitoring of bounding-box instability, feature-distribution anomaly detection, and ensemble-based detection strategies. To simulate realistic deployment scenarios, researchers collected a controlled dataset of both clean and patched objects under various environmental conditions. The detection performance is evaluated using precision, recall, F1-score, mean Average Precision (mAP), and inference latency.

Early investigation of research studies suggests that the use of ensemble detection, along with tracking confidence distributions, significantly improves the detection rates of adversarial patch attacks while keeping performance near real-time levels. This research focuses on identifying, rather than stopping, physical adversarial patch attacks using object detection-based defense strategies.