Adversarial Patch: Autonomous Vehicles

Adversarial patch applied to the safety and security of autonomous vehicles

Erick Constant
Student
Hampton University
erick.constant@my.hamptonu.edu

Chutima Boonthum-Denecke
Professor
Hampton University
Chutima.boonthum@hamptonu.edu

ABSTRACT

When it comes to the safety of autonomous vehicles using computer vision, we must first analysis the impact and risk that adversarial patches may present to its occupants, other drivers and property on the road. By seeing the result of the patch with the Ultralytics YOLO 11 model trained on things that an autonomous vehicle might encounter on the road we can perform an analysis of what could've been the impact without needing to run a simulation.

CCS CONCEPT

Computing methodologies

- Artificial intelligence
- Computer vision
- Computer vision tasks

KEYWORDS

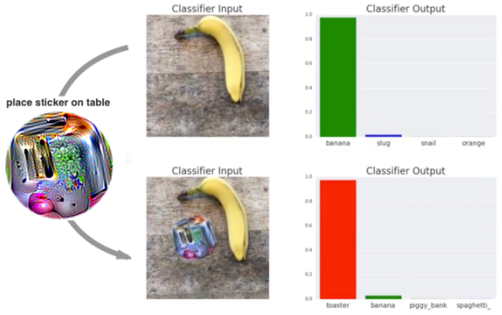
- Computer Vision
- Adversarial
- Security
- Self-Driving

Introduction:

Currently there has been an increase in automotive companies trying to deliver autonomous self-driving to their vehicles while also not dramatically increasing the cost of repair and the cost of the vehicle. To achieve autonomous self-driving, companies rely on cameras in combination with sensors to detect their environment and make movements based on those detections. Because of this, these vehicles heavily rely on artificial intelligence, we must explore the effect of adversarial attacks. This is because these applications affect the occupants of the vehicle, other vehicles on the road and property and pedestrians. By exploring this topic, we can see how a once glance computer vision model performs against adversarial attacks and apply reason to see how it might cause an injury when it comes to autonomous vehicles.

Related Work

In 2017, researchers from Google LLC published a paper called *Adversarial Patch* regarding a way to make adversarial image patches that can be deployed to attack computer vision models' accuracy in detection. Ultimately, they believe that it could "allows attackers to create a physical-world attack without prior knowledge of the lighting conditions, camera angle, type of classifier being attacked or even the other items within the scene.". This specifies that the research related to adversarial patches is relevant for any system that is reliant on computer vision models because of the fact these adversarial patches could be used anywhere to exploit these systems.



Example of a Adversarial Patch along with its corresponding result inside of the paper

Additionally, to add onto that research there have been papers published about attempting to find a way to generate new adversarial patches that have a higher success rate in disrupting artificial intelligence computer vision models such as what is listed in the paper called *AdvReal: Physical Adversarial Patch Generation Framework for Security Evaluation of Object Detection Systems.* Inside of this paper they found that their patch had a 90% average attack success rate due to their patches being both 2d and 3d.

Since then, there has been more research related to autonomous vehicles, mainly research related to the combination of multiple sensors and cameras called fusion systems. In a paper called *Exploring Adversarial Robustness of Multi-sensor Perception Systems in Self Driving*, it explores a similar concept and highlights how using image simulation, applying an adversarial patch could trick the camera to not recognize the vehicle. They believe that "successful attacks are primarily caused by easily corrupted image features".

Methodology

To start with this research, I first went online researching artificial intelligence computer visions models, because I previously had experience using the Ultralytics models, I went back to Ultralytics to compare the YOLO (You Only Look Once) models that were listed. During my research I found that YOLO11 was the better model due to its stability and speed compared to the older models of the YOLO market, and while the YOLO12 is released Ultralytics still recommends the YOLO11 for stability which is something a car company would've likely deployed than a recently released version.

After choosing the Ultralytics model, I then searched the internet for public datasets that had things an autonomous vehicle might encounter on the road, such as difference sizes of vehicles, bikes, motorcycles, pedestrians and different types of signage. Using this, I was able to then train the model using the Google Colab framework on the A100 NVIDIA GPU using the NVIDIA CUDA.

I then printed the adversarial patch that was attached to the paper *Adversarial Patch* as seen below:

To then accurately get a taste of the real-world environment, I went outside and found things that an artificial intelligence model might pickup and used a 12 MP camera that is included on the Apple iPhone. I then took pictures with the adversarial patch to keep a controlled environment. I chose not to worry about the day apart from the model due to the car being required to see both night and day to truly be autonomous.

After training the model 3 times, I then ran the photos I took onto the model and got like shown here and created a table listed below.

Stop sign with the adversarial patch with 2 distinct confidences of being a truck located where the stop sign is.

Stop sign without the adversarial patch with a 54% confidence that it's a stop

This highlights that over reliant on cameras for self-driving systems could be life-threatening because how easy an almost 8-year-old patch can still trick the modern computer vision models.

Based off the results, listed below in the table, we can see that the adversarial patch causes some incorrectness when it comes to detecting the object correctly. Applying this knowledge to a real-life scenario, this level of incorrectness could be fatal for any occupants of the vehicle, or other motor vehicles.

Patch Applied To:	Objects Detected	Original	Adversarial	Change Percentage
Car:	0/Stop Sign	0	0	0.00%
	Bus	0	0	0.00%
	Car	0.92	0	-92.00%
	Motorcycle	0	0	0.00%
	People	0	0	0.00%
	Truck	0	0.26	26.00%
Person:	0/Stop Sign	0	0	0.00%
	Bus	0	0	0.00%
	Car	0.79	0.79	0.00%
	Motorcycle	0	0	0.00%
	People	0	0	0.00%
	Truck	0	0.27	27.00%
Wall:	0/Stop Sign	0	0	0.00%
	Bus	0	0	0.00%
	Car	0.92	0.88	-4.00%
	Motorcycle	0	0	0.00%
	People	0	0	0.00%
	Truck	0	0	0.00%
Stop Sign:	0/Stop Sign	0.54	0	-54.00%
	Bus	0	0	0.00%
	Car	0	0	0.00%
	Motorcycle	0	0	0.00%
	People	0	0	0.00%
	Truck	0	0.5	50.00%

REFERENCES

Annotation. (2024, July). Autonomous vehicle dataset [Data set]. Roboflow Universe. https://universe.roboflow.com/annotation-2dedh/autonomous-vehicle-r4sxh

AsphaltWS. (2023, April). Stop sign detection dataset [Data set]. Roboflow Universe. https://universe.roboflow.com/asphaltws/stop-sign-detection-uv4u8

blusonik. (2024, May). Adversarial Patch on Imagenet Dataset. Roboflow Universe. Roboflow. Retrieved February 3, 2025, from https://universe.roboflow.com/blusonik/adversarial-patch-on-imagenet

Brown, T., Mane, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial Patch. arXiv preprint arXiv:1712.09665.

Carleton University. (2022, August). Yield dataset [Data set]. Roboflow Universe. https://universe.roboflow.com/carleton-ß university/yield-jvifi

Huang, Y., Ren, Y., Wang, J., Huo, L., Bai, X., Zhang, J., & Yu, H. (2025). AdvReal: Physical adversarial patch generation framework for security evaluation of object detection systems. arXiv. https://doi.org/10.48550/arXiv.2505.16402

Jocher, G., & Qiu, J. (2024). Ultralytics YOLO11 (Version 11.0.0) [Software]. Available from https://github.com/ultralytics/ultralytics

Toyproblem. (2025, July). Shapes_and_stop_sign dataset [Data set]. Roboflow Universe. https://universe.roboflow.com/toyproblem/shapes_and_stop_sign-eqddt

Trafficlight Predictor. (2025, December). Speed sign dataset [Data set]. Roboflow Universe. https://universe.roboflow.com/trafficlight-predictor/speed-sign-dmzmi

Tu, J., Li, H., Yan, X., Ren, M., Chen, Y., Liang, M., Bitar, E., Yumer, E., & Urtasun, R. (2021). Exploring adversarial robustness of multi-sensor perception systems in self driving. arXiv. https://doi.org/10.48550/arXiv.2101.06784