Talk Proposal: Towards the Realistic Evaluation of Evasion Attacks using CARLA

***Abstract*—In this talk we describe a content-preserving attack on object detectors, ShapeShifter [1], and demonstrate how to evaluate this threat in realistic scenarios. We describe how we use CARLA [2], a realistic urban driving simulator, to create these scenarios, and how we use ShapeShifter to generate content-preserving attacks against those scenarios.**

***Index Terms*—Adversarial attack, Object detection, CARLA**

I. Introduction

Most adversarial machine learning research focuses on indistinguishable perturbations while few focus on more realistic threats like content-preserving perturbations and non-suspicious inputs [3]. Although some research does focus on these more realistic threats [4], they are often limited to the classification task even though many safety- and security-critical systems need to localize, in addition to classify, objects. These systems often perceive the physical world with sensors, and most attacks do not account for the characteristics of this sensing pipeline and thus fail to remain adversarial when physically realized.

II. The ShapeShifter Attack

A recent work, ShapeShifter [1], creates content-preserving physically-realizable adversarial stop signs that are misdetected as other objects by the Faster R-CNN Inception-v2 object detector trained on the MS-COCO dataset. More recently, we showed how to make this attack an accessory [5] by limiting the perturbation to the printable area of a t-shirt (Fig. 1).

To create these attacks, we first had to digitally craft them, then physically fabricate them, and finally test them in the real world. This process is straightforward assuming one knows how to accurately model these real-world transformations that Expectation over Transformation [6] requires, a method ShapeShifter relies upon. More often there was a non-trivial amount of iteration between crafting, fabricating, and testing. Additionally, choosing how to perturb objects that are realistically and physically constrained (e.g., attackers cannot typically perturb the sky or even all parts of the object in non-suspicious ways) took some consideration. Our experience has led us to seek means to reduce these difficulties.

III. The CARLA Simulator

CARLA is an open-source realistic driving simulator. It provides a convenient way to train autonomous driving agents with a variety of different sensor suites (cameras, LIDAR, depth, etc.), and test them in realistic traffic scenarios. By integrating with CARLA, we reduce the time between digitally crafting a perturbation and testing it against real scenarios. In some cases, we have found our attacks already work in CARLA (Fig. 2). CARLA has also enabled us to craft new perturbations that were previously difficult to model (Fig. 3). The reproducibility of the CARLA scenarios enables us to better understand the efficacy of our attacks.

Our purpose in creating these attacks and evaluating them under realistic scenarios is to understand whether adversarial examples pose a real threat to real systems. While it is true that current adversaries might use easier means to subvert these systems (e.g., by knocking over a stop sign), we should not underestimate what methods attackers will undertake to achieve their means. We believe the creation of these realistic evaluations will increase our defensive capabilities. To date it is unclear whether the most successful defense, adversarial training [7], can even scale to ShapeShifter-style perturbations. A more recent defense has shown some defensive capability against ShapeShifter-like attacks on image classifiers [8]. However, neither of these defenses think about the system as a whole. By creating these realistic evaluations in CARLA, defenders can think more holistically about potential strategies. For example, CARLA enables us to quickly experiment with different sensing modes. We have found that the depth channel might provide some defensive capability against pixel perturbations when the object of interest has a distinct shape (Fig. 4). From a systems perspective, perhaps this is an acceptable defense, because now the attacker must now perturb the geometry of the object in addition to its color.

References

[1] S.-T. Chen, C. Cornelius, J. Martin, and D. H. P. Chau, "ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 52–68, Springer, 2018.

[2] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16, 2017.

[3] J. Gilmer, R. P. Adams, I. Goodfellow, D. Andersen, and G. E. Dahl, "Motivating the Rules of the Game for Adversarial Example Research," *arXiv e-prints*, Jul 2018, 1807.06732.

[4] X. Zeng, C. Liu, Y.-S. Wang, W. Qiu, L. Xie, Y.-W. Tai, C. Keung Tang, and A. L. Yuille, "Adversarial Attacks Beyond the Image Space," *arXiv e-prints*, Nov 2017, 1711.07183.

[5] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, (New York, NY, USA), pp. 1528–1540, ACM, 2016.

[6] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing Robust Adversarial Examples," *arXiv e-prints*, Jul 2017, 1707.07397.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in *International Conference on Learning Representations*, 2018.

[8] E. Chou, F. Tramèr, G. Pellegrino, and D. Boneh, "SentiNet: Detecting Physical Attacks Against Deep Learning Systems," *arXiv e-prints*, Dec 2018, 1812.00292.

Fig. 1. An accessorized [5] t-shirt created by the ShapeShifter attack [1] causes the object detector to see a bird rather than a person. Faces anonymized.

Fig. 2. Transferring the attack onto the CARLA pedestrian model. We found the perturbation remained adversarial for a limited set of camera positions and orientations.

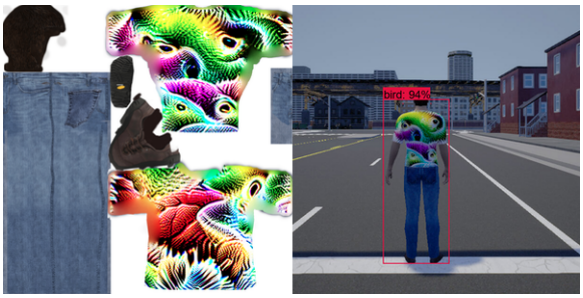

Fig. 3. CARLA enables us to perturb the environment with realistic constraints. Here we demonstrate a full t-shirt perturbation that remains adversarial from more camera positions and orientations.

Fig. 4. In autonomous driving systems, there is reason to believe adversarial examples like ours can be mitigated using multi-modal sensing. Here we show an off-the-shelf object detector seeing a stop sign using only the depth channel. CARLA enables us to quickly test and prototype these kind of defenses.