Demo: Sequential Attacks on Kalman Filter-based Forward Collision Warning Systems

Yuzhe Ma, Jon Sharp, Ruizhe Wang, Earlence Fernandes, Xiaojin Zhu
University of Wisconsin–Madison

***Abstract*—Kalman Filter (KF) is widely used in various domains to perform sequential learning or variable estimation. In the context of autonomous vehicles, KF constitutes the core component of many Advanced Driver Assistance Systems (ADAS), such as Forward Collision Warning (FCW). It tracks the states (distance, velocity etc.) of relevant traffic objects based on sensor measurements. The tracking output of KF is often fed into downstream logic to produce alerts, which will then be used by human drivers to make driving decisions in near-collision scenarios. In this work, we demonstrate planning-based attacks on Forward Collision Warning — a machine-human hybrid system that uses KF. Based on our work published at the AAAI-2021 conference, we use an MPC-based algorithm and show how an attacker can sequentially perturb vision measurements to change the FCW alert signals at desired points in time. We simulate our attack on CARLA using standard test protocols from the National Highway Traffic Safety Administration.**

I. Demo Description

Advanced Driver Assistance Systems (ADAS) are hybrid human-machine systems that are widely deployed on production passenger vehicles. They use sensing, traditional signal processing and machine learning to detect and raise alerts about unsafe road situations and rely on the human driver to take corrective actions. Popular ADAS examples include Forward Collision Warning (FCW), Adaptive Cruise Control and Autonomous Emergency Braking.

Although ADAS hybrid systems are designed to increase road safety when drivers are distracted, attackers can negate their benefits by strategically tampering with their behavior. For example, an attacker could convince an FCW or AEB system that there is no imminent collision until it is too late for a human driver to avoid the crash.

This demo is based on recent work (appearing at AAAI-2021) that studies the robustness of ADAS to attacks. The core of ADAS typically involves tracking the states (e.g., distance and velocity) of road objects using Kalman filter (KF). Downstream logic uses this tracking output to detect unsafe situations before they happen. We focus our efforts on Forward Collision Warning (FCW), a popular ADAS deployed on production vehicles today. FCW uses KF state predictions to detect whether the ego vehicle (vehicle employing the ADAS system) is about to collide with the most important object in front of it and will alert the human driver in a timely manner. Thus, our concrete attack goal is to trick the KF that FCW uses and make it output incorrect state predictions that would induce false or delayed alerts depending on the specific physical situation.

Recent work has examined the robustness of road object state tracking for autonomous vehicles [2]. Their attacks create an instantaneous manipulation to the Kalman filter inputs without considering its sequential nature, the downstream logic that depends on filter output, or the physical dynamics of involved vehicles. This leads to temporarily hijacked Kalman filter state predictions that are incapable of ensuring that downstream logic is reliably tricked into producing false alerts. By contrast, we adopt an online planning view of attacking KFs that accounts for: (1) their sequential nature where current predictions depend on past measurements; and (2) the downstream logic that uses KF output to produce warnings. Our attack technique also considers a simplified model of human reaction to manipulated FCW warning lights.

Based on our model predictive control algorithm published at AAAI-2021, we demonstrate attacks that can hijack FCW behavior. Our attacks force FCW alerts that mask the true nature of the physical situation involving the vehicles until it is too late for a distracted human driver to take corrective actions. Our demon uses a high-fidelity driving simulation using CARLA, a popular tool for autonomous vehicle research and development. We create test scenarios based on real-world driving data [1] and demonstrate the practicality of the attack in causing crashes involving the victim vehicle.

Specifically, we show that attack planning in advance of the targeted point is beneficial compared to without planning. Given 25 steps of planning (or 1.25 seconds based on specific physical situations in our tests) before the targeted time point, the attacker can cause the desired effect, while the attack fails without planning.

Acknowledgements. This work was supported in part by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. XZ acknowledges NSF grants 1545481, 1704117, 1836978, 2041428, 2023239 and MADLab AF CoE FA9550-18-1-0166.

References

[1] European New Car Assessment Programme, “Euro NCAP LSS Test Protocol. Version 2.0.1,” 2018. [Online]. Available: https://cdn.euroncap.com/media/26996/euro-ncap-aeb-c2c-test-protocol-v20.pdf

[2] Y. Jia, Y. Lu, J. Shen, Q. A. Chen, H. Chen, Z. Zhong, and T. Wei, “Fooling detection alone is not enough: Adversarial attack against multiple object tracking,” *2020 International Conference on Learning Representations (ICLR)*, 2020.

Workshop on Automotive and Autonomous Vehicle Security (AutoSec) 2021
25 February 2021, Virtual
ISBN 1-891562-68-1
https://dx.doi.org/10.14722/autosec.2021.23015
www.ndss-symposium.org