Improving Security and User Privacy in Learning-Based Traffic Signal Controllers (TSC)

Center for Transportation, Environment, and Community Health
Final Report

by
Ammar Haydari, Michael Z. Zhang, Chen-Nee Chuah

February 28, 2022

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No.	2.Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Improving Security and User Privacy in Learning-Based Traffic Signal Controllers (TSC)	5. Report Date March 31, 2022
	6. Performing Organization Code
7. Author(s) Chen-Nee Chuah (ORCID ID #: 0000-0002-2772-387X)	8. Performing Organization Report No.
9. Performing Organization Name and Address Department of Electrical & Computer Engineering University of California, Davis Davis, CA 95618	10. Work Unit No.
	11. Contract or Grant No. 69A3551747119
12. Sponsoring Agency Name and Address U.S. Department of Transportation 1200 New Jersey Avenue, SE Washington, DC 20590	13. Type of Report and Period Covered Final Report 04/01/2021 – 03/31/2022
	14. Sponsoring Agency Code US-DOT

15. Supplementary Notes

16. Abstract

The 21st century of transportation systems leverages intelligent learning agents and data-centric approaches to analyze information gathered with sensing (both vehicles and roadsides) or shared by users to improve transportation efficiency and safety. Numerous machine learning (ML) models have been incorporated to make control decisions (e.g., traffic light control schedules) based on mining mobility data sets and real-time input from vehicles via vehicle-to-vehicle and vehicle-to-infrastructure communications. However, in such situations, where ML models are used for automation by leveraging external inputs, the associated security and privacy issues start to surface. This project aims to study the security of ML systems and data privacy associated with learning-based traffic signal controllers (TSCs). Preliminary work has demonstrated that deep reinforcement learning (DRL) based TSCs are vulnerable to both white-box and black-box cyber-attacks. Research goals include 1) quantifying the impact of such security vulnerabilities on the safety and efficiency of the TSC operation, and 2) developing effective detection and mitigation mechanisms for such attacks. In learning based TSCs, vehicles share their messages with the DRL agents at TSCs, which will then analyze the data and take action. Sharing vehicular mobility data with a network of TSCs may cause privacy leakage. To address this problem, differential privacy techniques will be applied to the mobility datasets to protect user privacy while preserving the effectiveness of the prediction outcomes of traffic-actuated or learning-based TSC algorithms. Approaches will be evaluated in vehicular simulators using real mobility data from San Francisco and other cities in California. By accomplishing these goals, learning-based transportation systems will be more secure and reliable for real-time implementations.

17. Key Words Deep reinforcement learning, traffic signal control, adversarial attack, security defense, statistical anomaly detection	18. Distribution Statement Public Access Accepted for publication in IEEE Open Journal of Intelligent Transportation Systems

19. Security Classif (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No of Pages	22. Price

Final Project Report

Summary of Accomplishments:
Chen-Nee Chuah, Michael Zhang, and their PhD student, Ammar Haydari, studied the impact of deep reinforcement learning (DRL)-based traffic signal controller (TSC) on air quality using real traffic demands on city-level road networks. We studied a major DRL approach called advantage actor-critic (A2C) using multi-agent settings on a synthetic multi-intersection network and on a real traffic network of San Francisco downtown with 24 hours traffic dataset. Our results indicate that learning based DRL methods achieved the lowest air pollution level on synthetic networks even with a simple delay-based reward function [1]. The team also studied the vulnerabilities of these DRL-TSC algorithms in the presence of black-box and white-box adversarial attacks. Our results show that the performance of DRL learning agent decreases in both settings, resulting in higher levels of traffic congestion. We then proposed an ensemble model to perform sequential anomaly detection of the adversarial attacks. Our model minimizes detection delay, achieves lower false alarm rates due to cumulative anomaly inspection [2].

1. Introduction

Intelligent transport systems (ITS) integrate information and communication technologies (ICT) with transportation applications that increase traffic efficiency and security for all participants, such as pedestrians and vehicles. The latest technological improvements increased the quality of transportation. New data-driven approaches bring out a new research direction for all control-based systems, e.g., in transportation, robotics, IoT and power systems. Combining data-driven applications with transportation systems plays a key role in recent transportation applications. Deep reinforcement learning (DRL) is conceived to increase the traffic efficiency in ITS by enabling a learning structure that interacts with the environment. While many DRL-based traffic optimization methods are presented for different ITS applications, the majority of these applications are concentrated on TSCs. DRL-TSCs have the potential to offer a solution by decreasing the travel delay and increasing the traffic efficiency. Our performance analysis regarding the applicability of DRL-TSC, other studies in the literature, and several industrial efforts demonstrate that data-driven ITS controllers are expected to occupy the roads in near future

To date, there remains a limited understanding of the security vulnerabilities of learning based ITS controllers and their impact on various operational performance metrics. In our project, we experimented another research direction of ITS security where we characterize the security vulnerabilities of TSCs when implemented with DRL model and proposed a novel statistical detection model. Main-stream adversarial attack models continuously inject adversarial samples to the learning models and expect to fool the model quickly. To protect the DRL-TSC learning model we propose to use statistical sequential detection models with a novel ensemble detection algorithm that achieves the best detection performance in all cases.

2. Adversarial Attacks on DRL-TSCs

Controlling ITS components with a learning-based model opens a new attack surface for adversaries[3]. Misleading the behavior of ITS controllers with adversarial samples may result in life-threatening conditions. One of the main application areas of learning-based controller models is TSC with DRL. Therefore, the security analysis of DRL-TSCs needs to be investigated. We identified two main threat models for DRL-TSCs: directly injecting minimal random perturbations to the learning controller or sending falsified information to the TSC using Sybil or compromised vehicles. The security breach of DRL-TSC in the case of these attack surfaces is a challenge.

In this project, we thoroughly study security vulnerabilities of DRL-based TSCs under two gradient-based adversarial attack models namely Fast Gradient Sign Method (FGSM)[4] and Jacobian-based Saliency Map Attack (JSMA)[5] with white-box and black-box settings. Since gradient-based adversarial attacks such as FGSM and JSMA generally have a minimal perturbation on the data, it is also hard to differentiate adversarial samples from real samples with standard anomaly detectors. The threat model of adversarial attacks on DRL-TSCs is shown in Figure 1.

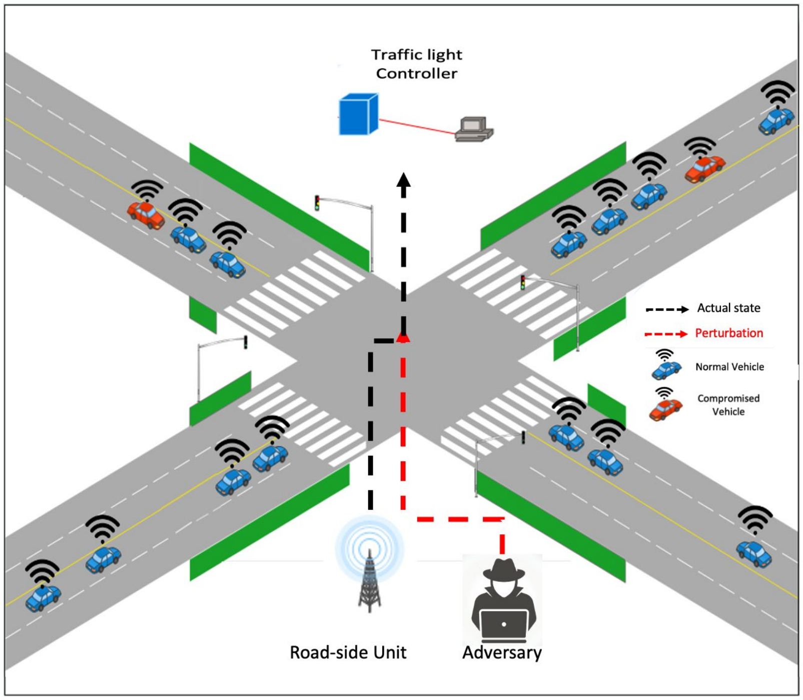

Figure 1. TSC is controlled with a DRL agent. An adversary can attack the agent with falsified data, disturbing the input state.

Given the adversarial attacks FGSM and JSMA for single intersection and multi-intersection scenarios, we proposed a statistical anomaly detector to detect even infinitesimally small anomalies. An ensemble anomaly detector that combines two sequential anomaly detection models and an autoencoder-based anomaly detection model with a CUSUM-like detection model is evaluated on the gradient-based adversarial attacks. The experiments show that the proposed ensemble sequential anomaly detection model achieves the best detection rate with different DRL agents and TSC scenarios.

3. Adversarial Attack Performance

In the white-box attack model, the adversary launches the attacks on DRL-based TSCs by injecting anomalies to the original input state. Since DNN is the policy of a learning agent, selecting the correct action of the DRL agent will be affected by the white-box attack. In the black-box attack scenario, the attacker does not have a precise knowledge about the model. We investigate the vulnerability of the DNN policies for DRL-based TSCs when the attacker does not have access to the actual target model.

For FGSM attack, an attacker will perturb the input state with very small changes that are invisible by the controller. As pointed out in the original FGSM paper [2], minimal perturbation leads to the DNN classifying output to a wrong class.

For JSMA attack, the attacker constructs the saliency map of given input state with respect to randomly selected action using the forward gradient of the DNN. In this attack model, we found that the attacker needs to perturb at least 40% of the feature dimensions to mislead the DRL agent, hence, we selected perturbation parameters accordingly.

References

[1] Chia-Cheng Yen, Dipak Ghosal, Michael Zhang, and Chen-Nee Chuah, "Security vulnerabilities and protection algorithms for backpressure-based traffic signal control at an isolated intersection." *IEEE Transactions on Intelligent Transportation Systems* (2021).

[2] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

[3] Papernot, Nicolas, et al. "The limitations of deep learning in adversarial settings." *2016 IEEE European symposium on security and privacy (EuroS&P). IEEE*, 2016.

[4] A. Haydari, H. M. Zhang, C-N. Chuah, and D. Ghosal, "Impact of Deep RL-based Traffic Signal Control on Air Quality, *IEEE Vehicular Technology Conference (VTC 2021)-Spring*, April 2021.

[5] A. Haydari, M. Zhang, and C-N. Chuah, "Adversarial Attacks and Defense in Deep Reinforcement Learning (Deep-RL) Based Traffic Light Controller, submitted to *IEEE Open Journal of Intelligent Transportation Systems.*