Thesis Project Portfolio

Structured Interpretable Manipulation of Policies

(Technical Report)

Analyzing Wariness of Autonomous Vehicles Through the Vehicle Decision-Making Process

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Quinlan Dawkins

Spring, 2020

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

Recommender systems, medical imaging, and vehicle automation are all examples of modern problems that lean on artificial intelligence techniques for solutions. Artificial intelligence, especially when machine learning (ML) techniques are employed, are unique in that their outputs can be difficult to interpret. The research presented in this portfolio explores interpretability and how it limits access to modern systems. From a technical perspective, consider deep neural networks (DNNs), which are used to approximate solutions to highly complex problems. With increasingly complex problems and models, recent studies have shown strange behavior of DNNs in image recognition tasks when imperceptible changes are made to an image. Because machine learning algorithms rely on being able to infer behavior from data, when data is either tampered with or new to the model, behavior can be unpredictable. Tampering with the input to a data driven algorithm is referred to as an adversarial attack and poses a potential risk when relying on machine learning models to drive important systems. Additionally, adversarial examples do not have to come from tampered data and can occur in more natural data (Zhao, 2018). From a social perspective, the inability to interpret the behavior of a system prevents potential actors from joining a space that relies on ML solutions. Here, an example of this is analyzed in autonomous vehicles.

The technical thesis explores the interpretability of adversarial attacks in reinforcement learning (RL) environments. A key feature of adversarial attacks pertains to the limited magnitude of an attack being able to have a significant impact on the output of the target. This is closely tied to model interpretability which involves understanding model outputs and why such a small change in the input is able to completely change the output of a DNN (Tao, Ma, Liu, &

Zhang, 2018). This naturally leads to a question of interpretability of adversarial attacks. In this research, we explore a technique used in a classification setting for improving group sparsity of adversarial attacks in an RL setting. Additionally we adapt methods used to measure attack interpretability in a classification setting to the RL setting where the objective is to maximize a cumulative reward. Future work could explore other RL models or defense tactics.

The STS thesis analyzes how consumers, investors, and regulators remain wary of entering the space of autonomous vehicles (AVs) by using standards analysis. Specifically, the problem of designing a vehicle decision-making process is the major technological challenge and point of controversy with AVs. Vehicle decision-making systems can be broken down into a sensing system, client system, action system, and human-machine interface (Martínez-Díaz & Soriguera, 2018). First, basic ethical, legal, and environmental challenges are described to set up the analysis of each part of the decision-making system using standards. Standards analysis is designed to simplify complex systems. In this case, it is a powerful tool that can improve the ability to reason with diversity in the space of AVs. The analysis in this paper then serves to aid in overcoming the wariness of potential actors in the space.