Andrew Garcia

4/1/21

ECE Graduate Seminar Spring 2022

Trustworthy Reinforcement Learning for Safety-Critical Systems

This report is a summary for a seminar lecture given on March 25th, 2022 by Virginia Tech Professor Dr. Ming Jin. The topic of this lecture is "Trustworthy Reinforcement Learning for Safety-Critical Systems" which involves two things, Reinforcement Learning and Safety-Critical Systems. Reinforcement learning is a subfield of machine learning that uses mathematical models and algorithms to learn from experience. Safety critical systems are those that have a high degree of risk, and the failure of which could cause loss or injury to people. Examples include industrial control systems, aircraft flight controls, nuclear power plants and medical devices. In this report we will explore the importance of combining these two concepts.

The goal of Dr. Ming Jin's research is to "Develop theories of optimization, control, and machine learning for the next-generation trustworthy AI." This means that this model should be robust, data efficient, explainable, and adaptable. Containing these four characteristics makes for a resilient Reinforcement Learning model. Humans will always be in the loop, as Dr. Jin states, "whether we like it or not," which is the main reason why these models must be human centric.

Reinforcement Learning is a branch of machine learning that deals with the problem of maximizing an agent's cumulative reward function, which is a function that measures how well

an action performed by the agent will increase its own future rewards. In other words, Reinforcement Learning aims to optimize decision making to maximize expected utility and minimize cost. The term “trustworthy” is used to describe a reinforcement learning algorithm that can be trusted to make the right decision in every situation. In other words, we want an RL algorithm that will always learn what it should do in every situation, without making mistakes or taking unnecessary actions.

One example for testing models is the City-Learn Challenge. From their website, the City-Learn Challenge is “an opportunity for researchers from multi-disciplinary fields to investigate the potential of artificial intelligence and distributed control systems to tackle multiple problems within the energy domain.” The basic idea of this challenge is, you have a lot of buildings in the community, and you want to know how to optimize the storage in a distributed fashion to coordinate them to support a grid.

Reinforcement Learning has many applications; however, it has great potential for creating trustworthy systems to make our cities and modern life a better place to live in. It is not only difficult to create a RL model, but it is even more difficult to create once that can handle the complexities of our city’s infrastructure. Achieving goals like the ones set out in this report will lead to a better and safer world for everyone.