Specification-Guided Reinforcement Learning

Suguman Bansal
School of Computer Science, Georgia Institute of Technology

Overview

Unprecedented advances in AI have led to AI-based devices making complex decisions in our daily lives, with *Reinforcement Learning* (RL) increasingly deployed for consequential decisions in areas like autonomous cars, chatbots, and medical diagnosis. However, the black-box nature of modern RL-enabled systems has made issues of **trust** central in AI research today. The mistrust is aggravated further as tasks in RL are usually specified in the form of rewards that are non-interpretable and non-verifiable.

The PI's CAREER goal is to address the pressing issue of trust in RL-enabled systems by integrating *Formal Methods* (FM) into RL. The proposal revolves around two fundamental pillars. Firstly, we advocate representing desired tasks using *temporal logic specifications*, which offer interpretability, rigor, and verifiability, in contrast to conventional reward functions. Yet another advantage of temporal specifications lies in their availability before exploration, enabling the design of specification-aware RL algorithms. Secondly, departing from rigorous notions of correctness in FM, we explore notions of trustworthiness tailored to the RL community, thus acknowledging the uncertainties and unknowns prevalent in RL settings. This proposal defines trustworthy behavior in RL through compositionality, generalizability, and verifiability.

Key Words Reinforcement Learning, Temporal Logic, Formal Methods, Trustworthiness

Intellectual Merit

This proposal offers a new perspective of trustworthy RL based on Formal Methods. It charts an agenda to leverage temporal logic specifications, as opposed to rewards, to improve the compositionality, generalizability, and verifiability of RL algorithms. The widespread consensus among researchers and technologists is that the prior works in the design of RL-enabled systems lack a methodological approach. In contrast, this proposal will unify these problems under the framework of temporal logics and will present principled developments via the combination of symbolic information from the logical formalism with low-level reasoning available from RL. This proposal focuses on laying theoretical foundations, algorithmic and tool development, and the creation of benchmarks. The proposal will necessitate the design of novel sampling-based techniques and algorithms, as the differences in environment assumptions prevent lifting works from FM into RL.

Broader Impacts

The broad societal impact of trustworthy AI cannot be understated. As evidenced by articles in popular news media over the past few years, there is overwhelming evidence that AI systems are brittle and could lead to catastrophic failures costing not only large capital but also human lives. In this light, the success of our proposal in building foundations of trustworthy RL could revolutionize the way we design all software and hardware. Trustworthy RL could be deployed for system development in safety-critical and security-critical applications, including robotics and automation, medical diagnostics, and banking.

From a technical standpoint, the proposal addresses challenges from RL using formal methods techniques. Problems considered here could have a broader impact beyond RL. For instance, the formalism of generalization that we develop for RL could be used to improve generalization in other sequential decision-making applications. Hence, this proposal could introduce novel problems to the formal methods community. The proposal research also drives an interdisciplinary effort across multiple areas of CS, including FM, AI, deep learning, and theory of computation. This presents a rich opportunity to develop new pedagogical materials for undergraduate and graduate researchers.