Assured Reinforcement Learning

Prof.Suresh Jagannathan
Purdue University

Abstract of the talk :

Despite the tremendous advances that have been made in the last decade on developing useful machine-learning applications, their wider adoption has been hindered by the lack of strong assurance guarantees that can be made about their behaviour. In this talk, we consider how formal verification techniques developed for traditional software systems can be repurposed to ensure the safety of reinforcement learning (RL)-enabled ones, a particularly important class of machine learning systems for which assurance guarantees are especially critical. We consider two complementary approaches.

The first defines a black box verification toolchain that reasons about correctness extensionally, using a syntax-guided inductive synthesis framework that generates a simpler and more malleable deterministic program guaranteed to represent a safe control policy of an RL system. Our synthesis procedure is designed with verification in mind and is thus structured to incorporate formal safety constraints drawn from a logical specification of the control system the network purports to implement, along with additional salient environment properties relevant to the deployment context. Rather than repairing the network directly to satisfy safety constraints, we instead treat the synthesized program as a safety shield that operates in tandem with the network, overriding network-proposed actions whenever such actions can be shown to lead to potentially unsafe states. Our pipeline thus retains performance, provided by the neural policy, while maintaining safety, provided by the program.

The second defines a white box correct-by-construction methodology that integrates an optimization-based abstraction refinement loop into the learning process. Our approach enables training to take place over an abstraction of a concrete network that operates over dynamically constructed partitions of the input space. Classical gradient descent methods used to optimize these networks can be seamlessly adapted to this framework to ensure the soundness of our approach. Notably, empirical results show that we can realize safety without compromising on accuracy, giving us a meaningful pathway for seamlessly integrating safety and precision within the training process.

This is joint work with He Zhu, Xuankang Lin, and Roopsha Samanta.