Safety Assurance for Systems with Machine Learning Components

Chelsea Sidrane
Stanford University
csidrane@stanford.edu

Abstract

The use of machine learning components in safety-critical systems creates reliability concerns. My thesis focuses on developing algorithms to address these concerns. Because the assurance of a safety-critical system generally requires multiple types of validation, my research takes three directions: safe deep learning algorithms, formal verification of neural networks, and adaptive testing methods.

Deep learning has proven useful in domains as diverse as computer vision, natural language processing, and control of autonomous agents. Given its success, there is interest in applying deep learning to safety-critical systems (Bojarski et al. 2016). Consequently, methods must be developed to assure the safety of systems containing neural networks. My thesis focuses on addressing this need through safe deep learning methods, formal verification of safety properties for neural networks, and adaptive testing strategies.

Research Question 1: **How does the use of human experts affect the safety and performance of imitation learning algorithms?** In the imitation learning paradigm, an autonomous agent is trained by an expert. This work focuses on a commonly used imitation learning framework known as DAgger (Ross, Gordon, and Bagnell 2011), and considers the use of human experts within the DAgger framework. Together with others in the lab, I tested various imitation learning algorithms for driving a real car. A key problem with using humans as experts in DAgger is that the DAgger algorithm does not provide the human expert with enough control authority, as DAgger frequently switches control between the agent, called the novice, and the expert. This procedure can result in unsafe oscillations when the human expert overcorrects in response to their reduced control authority. In addition, handing control over to a partially trained novice is dangerous in and of itself. Relevant literature focuses on minimizing the number of times that a human expert is queried during a robot learning task, but does not address safety issues that can arise during learning (Laskey et al. 2016). We build on recent work that uses Bayesian deep learning to generate a "confidence" for a trained policy, but which does not provide an approach for selecting a meaningful threshold on this confidence (Menda, Driggs-Campbell, and Kochenderfer 2018). To address issues of safety both during and after learning, we developed the Human-Gate DAgger (HG-DAgger) algorithm (Kelly et al. 2019). HG-DAgger uses Bayesian deep imitation learning and gives complete control authority to the human expert when they deem the situation unsafe. The average confidence value of the policy at the time of "handoff" is used to place a safety threshold on the confidence once training is complete. In effect, the human expert's judgement is used to determine the threshold separating safe and unsafe behavior. This calibrated confidence metric can be used to monitor the safety of the learned policy.

Research Question 2: **How can formal methods be used to verify the safety of systems that contain neural network control policies?** An example of such as system is an autonomous aircraft that uses a neural network as a flight controller. Recent work has made the formal verification of neural networks tractable (Liu et al. 2019). My work builds upon that progress to address the "closed loop system" – a dynamical system with a neural network control policy. Specifically, this work addresses discrete time, nonlinear dynamical systems. There is recent literature from the hybrid systems community that studies the formal verification of continuous time dynamical systems with neural network control policies (Xiang et al. 2018). However, this work cannot be applied to discrete time systems without significant degradation in effectiveness. There is also work in the formal methods literature on verifying properties of nonlinear systems, but these methods cannot handle the computational burden of neural networks (Gao, Kong, and Clarke 2013).

Consequently, I have developed a method called OVERT to analyze such nonlinear, discrete time, closed loop systems with neural network control policies (Sidrane and Kochenderfer 2019). OVERT approximates the original nonlinear dynamical system in such a way that if no counter examples can be found for the approximate system, one can soundly claim there are no counter examples for the original system. OVERT produces tighter approximations than adapted continuous time tools and OVERT is much more computationally efficient than verification tools that are not specialized to handle neural networks (Gao, Kong, and Clarke 2013).

This work on OVERT will be extended in two directions over the next year. The first extension is to prove unbounded properties of discrete time systems, such as stability in the

control theory sense. My approach for this extension is to pursue Lyapunov stability, which requires finding a pair of timesteps such that the state space reachable at time $t + n$ is a subset of the state space reachable at time t for some finite n. I plan to use OVERT to search for such pairs of timesteps using iterative refinement of the initial set at time t and of the approximations used inside OVERT. The second extension of OVERT is to approximate not only the dynamical system, but also the neural network controller. The central idea is to approximate the neural network controller with a smaller network that is faster to verify. Once a smaller network has been trained or pruned, neural network verification tools can be used to confirm that the smaller, approximate network can be used to make sound claims about the original network.

Research Question 3: **When using adaptive stress testing, how do we know when a system had been tested *enough*?** Unlike formal methods, testing cannot prove the absence of bugs, but it can handle more complex systems than can formal verification tools. A recent technique called adaptive stress testing uses an adversarial agent trained using reinforcement learning to choose environment parameters that make the system under test fail (Lee et al. 2015). It is fast and flexible to solve this reinforcement learning problem using deep learning (Koren et al. 2018). Unfortunately, when deep learning is used, one cannot be guaranteed to find the optimal adversarial agent policy and corresponding "worst case scenario" for the original system. My current work is exploring whether useful notions of "testing coverage" can be obtained when doing adaptive stress testing. It would be useful to know where in the state space of the original system we've tested "enough". My first line of exploration uses deep, Bayesian, value-based reinforcement learning to train the adversarial agent (Lee et al. 2020). A Bayesian approach allows assessment of the confidence of the adversarial policy. High confidence of the adversarial policy should imply that enough samples have been taken from the region of state space in question. Work is ongoing to assess the reliability of these confidence estimates.

Research Plan By December 2020: Extend OVERT workshop paper for journal submission, begin exploring further closed loop verification topics. By February 2020: Prepare and submit paper on adaptive stress testing coverage work. By April 2021: Prepare and submit paper on unbounded closed loop verification technique. By June 2021: Prepare and submit paper on closed loop verification with approximation of the neural network. By August 2021: Finish first draft of thesis.

Anticipated Thesis Contributions My contributions consist of techniques to assure the safety of systems containing machine learning components using tools from a variety of fields including formal methods, testing, and Bayesian deep learning. Specifically, these contributions are: 1) A new imitation learning algorithm suitable for use with human experts that produces a heuristic estimate of policy safety once trained, 2&3) New algorithms with which to perform both bounded-time and unbounded time safety verification of closed-loop systems with neural network controllers, 4) A new algorithm with which to perform approximation of neural network controllers so that verification of closed-loop neural network systems may be done more efficiently, and finally 5) A method to assess testing coverage after adaptive stress testing has been used to find system failures.

References

Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* .

Gao, S.; Kong, S.; and Clarke, E. M. 2013. dReal: An SMT solver for nonlinear theories over the reals. In *International Conference on Automated Deduction*, 208–214. Springer.

Kelly, M.; Sidrane, C.; Driggs-Campbell, K.; and Kochenderfer, M. J. 2019. Hg-dagger: Interactive imitation learning with human experts. In *International Conference on Robotics and Automation (ICRA)*, 8077–8083. IEEE.

Koren, M.; Alsaif, S.; Lee, R.; and Kochenderfer, M. J. 2018. Adaptive stress testing for autonomous vehicles. In *IEEE Intelligent Vehicles Symposium (IV)*, 1–7. IEEE.

Laskey, M.; Staszak, S.; Hsieh, W. Y.-S.; Mahler, J.; Pokorny, F. T.; Dragan, A. D.; and Goldberg, K. 2016. Shiv: Reducing supervisor burden in DAgger using support vectors for efficient learning from demonstrations in high dimensional state spaces. In *IEEE International Conference on Robotics and Automation (ICRA)*, 462–469. IEEE.

Lee, K.; Laskin, M.; Srinivas, A.; and Abbeel, P. 2020. SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. *arXiv preprint arXiv:2007.04938* .

Lee, R.; Kochenderfer, M. J.; Mengshoel, O. J.; Brat, G. P.; and Owen, M. P. 2015. Adaptive stress testing of airborne collision avoidance systems. In *IEEE/AIAA Digital Avionics Systems Conference (DASC)*, 6C2–1. IEEE.

Liu, C.; Arnon, T.; Lazarus, C.; Barrett, C.; and Kochenderfer, M. J. 2019. Algorithms for verifying deep neural networks. *arXiv preprint arXiv:1903.06758* .

Menda, K.; Driggs-Campbell, K.; and Kochenderfer, M. J. 2018. EnsembleDAgger: A Bayesian approach to safe imitation learning. *arXiv preprint arXiv:1807.08364* .

Ross, S.; Gordon, G.; and Bagnell, D. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 627–635.

Sidrane, C.; and Kochenderfer, M. J. 2019. OVERT: Verification of Nonlinear Dynamical Systems with Neural Network Controllers via Overapproximation. *Safe Machine Learning Workshop at the International Conference on Learning Representations* .

Xiang, W.; Tran, H.-D.; Rosenfeld, J. A.; and Johnson, T. T. 2018. Reachable set estimation and safety verification for piecewise linear systems with neural network controllers. In *Annual American Control Conference (ACC)*, 1574–1579. IEEE.