Towards Formally Verified Deep Neural Networks

Submitted by Katie Dey on Sat, 03/10/2018 - 1:59pm. Contributors: Clark Barrett Guy Katz

BIO
Clark Barrett joined Stanford University in September 2016 after 14 years at NYU. His PhD dissertation (Stanford, 2003) introduced a novel approach to constraint solving called Satisfiability Modulo Theories (SMT). His current work focuses on the application of SMT solvers to improve reliability and security of software, hardware, and machine learning systems. He was also a pioneer in industrial applications of formal hardware verification, contributing to efforts at Intel and 0-in Design Automation. He has received best paper awards from DAC, ITC, and FMCAD, and is an ACM Distinguished Scientist.

ABSTRACT
Designing software controllers for autonomous systems is a difficult and error-prone task. On one hand, these systems are required to handle a wide variety of situations and corner cases; but on the other hand, they must meet a very high bar of capability, efficiency, responsiveness and reliability. In recent years, machine-learning (ML) has proven remarkably useful in addressing this challenge, and ML-generated artifacts are expected to become widespread in, e.g., autonomous vehicles. However, this trend has given rise to a new challenge: ML artifacts are, by nature, opaque, and are not amenable to traditional verification and certification. The present inability to automatically reason about these artifacts is likely to severely hinder their applicability to real-world situations.

In this talk we will survey our recent work on verifying deep neural networks, which are a state-of-the-art ML artifacts. Modern neural networks are known to perform well in general, but may present undesirable behavior in some situations. Neural network verification is the systematic exploration of the neural network in order to ensure that certain properties hold for any possible input that the network might encounter when deployed. This problem is computationally difficult (NP-complete); however, by bringing together techniques from the fields of linear programming and satisfiability modulo theories (SMT), we show how it is possible to greatly simplify the problem it in practice. Intuitively, this is done by deferring treatment of the non-linear elements of the problem (which are the main source of complexity) for as long as possible, and only considering them if/when needed. Using an initial implementation of this technique we were able to verify properties of a family of neural networks that are part of the ACAS Xu system: the next-generation airborne collision avoidance system, currently being developed by the FAA. We will also discuss how verification can be applied to ensure that a neural network is robust to small perturbations of their inputs ("adversarial robustness")

This talk is based on joint work with David Dill, Kyle Julian and Mykel Kochenderfer [1, 2]. This work was supported by Intel and by the FAA.


Other available formats:
Towards Formally Verified Deep Neural Networks
Switch to normal viewerSwitch to experimental viewer