Machine Learning (2023) 112:2787–2789
https://doi.org/10.1007/s10994-021-06113-4

Guest Editorial: Special issue on robust machine learning

Ransalu Senanayake[1] · Daniel J. Fremont[2] · Mykel J. Kochenderfer[3] · Alessio R. Lomuscio[4] · Dragos Margineantu[5] · Cheng Soon Ong[6]

Published online: 9 November 2021

This special issue of *Machine Learning* is devoted to exploring the emerging research questions in robust machine learning. Although machine learning techniques are currently deployed in various real-world systems, most of such applications are limited to low-risk systems such as search engines and recommender systems. However, recent advances in machine learning tools show their potential to be used in a variety of other applications, including safety-critical systems such as autonomous vehicles and high-cost industrial processes such as power plants. Deploying machine learning in such high-stake applications demands algorithms and tools to meet high robustness requirements.

Research efforts from multiple fronts are essential to confidently deploy machine learning in real-world systems. The machine learning models that we develop should take into account the uncertainty in safe regions of operation. Learned models should generalize to new phenomena and operate in previously unseen regions of the input space. In addition, there should be tools to analyze and verify whether a given system is indeed robust.

✉ Mykel J. Kochenderfer
mykel@stanford.edu

Ransalu Senanayake
ransalu@stanford.edu

Daniel J. Fremont
dfremont@ucsc.edu

Alessio R. Lomuscio
A.Lomuscio@imperial.ac.uk

Dragos Margineantu
dragos.d.margineantu@boeing.com

Cheng Soon Ong
chengsoon.ong@anu.edu.au

[1] Department of Aeronautics and Astronautics, Stanford University, 496 Lomita Mall, Durand Building, Room 227, Stanford, CA 94305, USA

[2] University of California, Santa Cruz 1156 High Street, MS SOE3, Santa Cruz, CA 95064, USA

[3] Department of Aeronautics and Astronautics, Stanford University, 496 Lomita Mall, Durand Building, Room 255, Stanford, CA 94305, USA

[4] Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK

[5] Boeing Research and Technology, P.O. Box 3707, M/C 50-201, Seattle, WA 98124, USA

[6] Data 61, CSIRO, Black Mountain Campus, Canberra, ACT 2601, Australia

Theoretical analyses help us establish confidence in the robustness of the system before deploying them in the real-world. Empirical studies can further validate that the system is robust.

Our call for papers welcomed both theoretical and empirical research studies on the robustness of machine learning. The editors received 43 papers and each paper was reviewed by at least 3 reviewers who are experts in the field. In total, 9 papers were selected for publication. In what follows, we summarize the accepted papers in this special issue.

In *Metrics and Methods for Robustness Evaluation of Neural Networks with Generative Models*, I. Buzhinsky, A. Nerinovsky, and S. Tripakis study the robustness of feed-forward deep neural networks in the presence of adversarial examples. The authors propose a framework and a set of metrics to measure robustness. They verify the suitability of these metrics on four image classification tasks. Because the paper focuses on "naturally plausible perturbations" to images, the analysis is easily extendable to many real-world applications.

Bayesian optimization has recently appeared as a successful tool for tuning hyperparameters in machine learning models. *Bayesian Optimization with Safety Constraints: Safe and Automatic Parameter Tuning in Robotics* by F. Berkenkamp, A. Krause, and A.P. Schoellig extends automatic parameter selection with safety in mind. This paper extends safe Bayesian Optimization to multiple safety constraints and applies the proposed theoretical framework to controlling a real quadcopter governed by a nonlinear controller.

Since the amount of data available to train models is growing rapidly, researchers often resort to distributed systems. However, the literature lacks a theoretical basis for understanding the robustness of such distributed training systems, for instance when some computers of the distributed system fail. *Byzantine-Robust Distributed Sparse Learning for M-Estimation* by J. Tu, W. Liu, and X. Mao presents a novel communication-efficient distributed algorithm to solve sparse M-estimation problems such as distributed LASSO.

M.D. Norton and J.O. Royset in their work titled *Diametrical Risk Minimization: Theory and Computations* propose a counterpart for Empirical Risk Minimization (ERM). The proposed method, unlike ERM, analyzes the "worst-case" empirical risk on a known set of training data. However, similar to ERM, the proposed method also has generalization bounds and can be easily implemented in practice, for instance with neural network classifiers.

Although support Vector Machines (SVMs) are a popular choice in machine learning, they are sensitive to outliers. *Unified SVM Algorithm Based on LS-DC Loss* by S. Zhou and W. Zhou unifies various SVM algorithms that are robust against such outliers. Unlike other methods aimed at solving nonconvex losses, the proposed model has a closed-form solution per iteration. The accuracy and the efficacy of the method are corroborated using benchmark regression and classification tasks.

In *Global Optimization of Objective Functions Represented by ReLU networks*, C.A. Strong, H. Wu, A. Zeljić, K.D. Julian, G. Katz, C. Barrett, and M.J. Kochenderfer study formal verification of neural networks. Specifically, the authors extend existing verifiers to provide provably optimal adversarial examples. The proposed approaches are benchmarked against a state-of-the-art solver on digit classification, aircraft collision avoidance, and aircraft localization datasets.

Neural networks are increasingly used in automated decision-making tasks. S. Katz, K. Julian, C. Strong, and M. Kochenderfer in *Generating Probabilistic Safety Guarantees for Neural Network Controllers*, introduces a method to determine the output properties that must hold for the controller to operate safely. Unlike previous methods, the proposed method uses stochastic dynamic models and provides probabilistic safety guarantees. The

method expresses model verification as a Markov Decision Process, which enables an estimate of the probability that a neural network reaches an unsafe state. The method is illustrated and discussed using the safety-critical application of aircraft collision avoidance.

In *Scenic: A Language for Scenario Specification and Data Generation*, D.J. Fremont, E. Kim, T. Dreossi, S. Ghosh, X. Yue, A.L. Sangiovanni-Vincentelli, and S.A. Seshia present a scenario description language for generating data according to a given description. Such tools are useful for stress testing and debugging various autonomous systems. For instance, Scenic has been applied to autonomous driving and Mars exploration simulations.

Even when neural networks are verified and stress tested, there is a chance that they will, intentionally or unintentionally, be adversarially attacked by corrupted data, unless the dataset is "sanitized." In *Stronger Data Poisoning Attacks Break Data Sanitization Defenses*, P.W. Koh, J. Steinhardt, and P. Liang propose three attack methods that can bypass common sanitization techniques. Such attacking tools help us develop even more robust defense mechanisms before we deploy the black-box models in the wild.

In summary, the papers that appear in this special issue consider different aspects of robust machine learning. They cover different ways to frame the question of robustness. The papers include tools to find safe hyperparameters, tools to analyze robustness, metrics to evaluate robustness, and methods to provide formal guarantees in machine learning-based systems. These tools will help enable us not only to operate autonomous systems safely and at low-cost but also help regulatory bodies assess the reliability of these systems. This special issue illustrates the rich class of problems to be considered and provides some initial forays into possible solutions. Further research efforts in these directions will help us use machine learning components in safety-critical real-world decision-making systems.

Acknowledgements The editors would like to thank all the authors for submitting their papers and revising them. We are indebted to the anonymous reviewers for providing thorough feedback on manuscripts within a short period of time.