Seminar Künstliche Intelligenz – Sommer-Semester 2021

Themen und Abstracts

Thema 1: DeepAbstract: Neural Network Abstraction for Accelerating Verification

Autoren: Pranav Ashok, Vahid Hashemi, Jan Kretınsky, und Stefanie Mohr.

Abstract. While abstraction is a classic tool of verification to scale it up, it is not used very often for verifying neural networks. However, it can help with the still open task of scaling existing algorithms to state-of-the-art network architectures. We introduce an abstraction framework applicable to fully-connected feed-forward neural networks based on clustering of neurons that behave similarly on some inputs. For the particular case of ReLU, we additionally provide error bounds incurred by the abstraction. We show how the abstraction reduces the size of the network, while preserving its accuracy, and how verification results on the abstract network can be transferred back to the original network.

23 Seiten, arXiv:2006.13735v1 [cs.LO] 24 Jun 2020

Ashok, P., Hashemi, V., Kretínský, J., and Mohr, S. (2020). Deepabstract: Neural network abstraction for accelerating verification. In Hung, D. V. and Sokolsky, O., editors, *Automated Technology for Verification and Analysis - 18th International Symposium, ATVA 2020, Hanoi, Vietnam, October 19-23, 2020, Proceedings*, volume 12302 of *Lecture Notes in Computer Science*, pages 92–107. Springer

Thema 2: Maximum Resilience of Artificial Neural Networks

Autiren: Chih-Hong Cheng, Georg Nührenberg, and Harald Ruess

Abstract. The deployment of Artificial Neural Networks (ANNs) in safety-critical applications poses a number of new verification and certification challenges. In particular, for ANN-enabled self-driving vehicles it is important to establish properties about the resilience of ANNs to noisy or even maliciously manipulated sensory input. We are addressing these challenges by defining resilience properties of ANN-based classifiers as the maximum amount of input or sensor perturbation which is still tolerated. This problem of computing maximum perturbation bounds for ANNs is then reduced to solving mixed integer optimization problems (MIP). A number of MIP encoding heuristics are developed for drastically reducing MIP-solver runtimes, and using parallelization of MIP-solvers results in an almost linear speed-up in the number (up to a certain limit) of computing cores in our experiments. We demonstrate the effectiveness and scalability of our approach by means of computing maximum resilience bounds for a number of ANN benchmark sets ranging from typical image recognition scenarios to the autonomous maneuvering of robots.

18 Seiten arXiv:1705.01040v2 [cs.LG] 5 Jul 2017

Cheng, C.-H., Nührenberg, G., and Ruess, H. (2017). Maximum resilience of artificial neural networks

Thema 3: Deep Save: A Data-driven Approach for Checking Adversarial Robustness in Neural Networks

Autoren: Divya Gopinath, Guy Katz, Corina S. Pasareanu, and Clark Barrett

Abstract. Deep neural networks have become widely used, obtaining remarkable results in domains such as computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, and bio-informatics, where they have produced results comparable to human experts. However, these networks can be easily "fooled" by adversarial perturbations: minimal changes to correctly-classified inputs, that cause the network to misclassify them. This phenomenon represents a concern for both safety and security, but it is currently unclear how to measure a network's robustness against such perturbations. Existing techniques are limited to checking robustness around a few individual input points, providing only very limited guarantees. We propose a novel approach for automatically identifying safe regions of the input space, within which the network is robust against adversarial perturbations. The approach is data-guided, relying on clustering to identify well-defined geometric regions as candidate safe regions. We then utilize verification techniques to confirm that these regions are safe or to provide counter-examples showing that they are not safe. We also introduce the notion of targeted robustness which, for a given target label and region, ensures that a NN does not map any input in the region to the target label. We evaluated our technique on the MNIST dataset and on a neural network implementation of a controller for the next-generation Airborne Collision Avoidance System for unmanned aircraft (ACAS Xu). For these networks, our approach identified multiple regions which were completely safe as well as some which were only safe for specific labels. It also discovered several adversarial perturbations of interest.

17 Seiten arXiv:1710.00486v1 2017

Gopinath, D., Katz, G., Pasareanu, C. S., and Barrett, C. W. (2018). Deepsafe: A data-driven approach for assessing robustness of neural networks. In Lahiri, S. K. and Wang, C., editors, *Automated Technology for Verification and Analysis - 16th International Symposium, ATVA 2018, Los Angeles, CA, USA, October 7-10, 2018, Proceedings*, volume 11138 of *Lecture Notes in Computer Science*, pages 3–19. Springer

Thema 4: The Marabou Framework for Verification and Analysis of Deep Neural Networks

Autoren: Guy Katz, Derek A. Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zelji , David L. Dill, Mykel J. Kochenderfer, and Clark Barrett

Abstract. Deep neural networks are revolutionizing the way complex systems are designed. Consequently, there is a pressing need for tools and techniques for network analysis and certification. To help in addressing that need, we present Marabou, a framework for verifying deep neural networks. Marabou is an SMT-based tool that can answer queries about a network's properties by transforming these queries into constraint satisfaction problems. It can accommodate networks with different activation functions and topologies, and it performs high-level reasoning on the network that can curtail the search space and improve performance. It also supports parallel execution to further enhance scalability. Marabou accepts multiple input formats, including protocol buffer files generated by the popular TensorFlow framework for neural networks. We describe the system architecture and main components, evaluate the technique and discuss ongoing work.

10 Seiten https://doi.org/10.1007/978-3-030-25540-4_26

Katz, G., Huang, D. A., Ibeling, D., Julian, K., Lazarus, C., Lim, R., Shah, P., Thakoor, S., Wu, H., Zeljic, A., Dill, D. L., Kochenderfer, M. J., and Barrett, C. W. (2019). The marabou framework for verification and analysis of deep neural networks. In Dillig, I. and Tasiran, S., editors, *Computer Aided Verification - 31st International Conference, CAV 2019, New York City, NY, USA, July 15-18, 2019, Proceedings, Part I*, volume 11561 of *Lecture Notes in Computer Science*, pages 443–452. Springer

Thema 5: Intriguing properties of neural networks

Autoren: Christian Szegedy, Dumitru Erhan, Wojciech Zaremba Ilya Sutskever, Ian Goodfellow, Joan Bruna, Rob Fergus

Abstract. Deep neural networks are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. While their expressiveness is the reason they succeed, it also causes them to learn uninterpretable solutions that could have counter-intuitive properties. In this paper we report two such properties. First, we find that there is no distinction between individual high level units and random linear combinations of high level units, according to various methods of unit analysis. It suggests that it is the space, rather than the individual units, that contains the semantic information in the high layers of neural networks. Second, we find that deep neural networks learn input-output mappings that are fairly discontinuous to a significant extent. We can cause the network to misclassify an image by applying a certain hardly perceptible perturbation, which is found by maximizing the network's prediction error. In addition, the specific nature of these perturbations is not a random artifact of learning: the same perturbation can cause a different network, that was trained on a different subset of the dataset, to misclassify the same input.

10 Seiten, arXiv:1312.6199v4

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks

Thema 6: Automated Verification of Neural Networks: Advances, Challenges and Perspectives

Autoren: Francesco Leofante, Nina Narodytska, Luca Pulina, Armando Tacchella

Abstract. Neural networks are one of the most investigated and widely used techniques in Machine Learning. In spite of their success, they still find limited application in safety- and security-related contexts, wherein assurance about networks' performances must be provided. In the recent past, automated reasoning techniques have been proposed by several researchers to close the gap between neural networks and applications requiring formal guarantees about their behavior. In this work, we propose a primer of such techniques and a comprehensive categorization of existing approaches for the automated verification of neural networks. A discussion about current limitations and directions for future investigation is provided to foster research on this topic at the crossroads of Machine Learning and Automated Reasoning.

7 Seiten arXiv:1805.09938 [cs.AI]

Leofante, F., Narodytska, N., Pulina, L., and Tacchella, A. (2018). Automated verification of neural networks: Advances, challenges and perspectives. *CoRR*, abs/1805.09938

Thema 7: Ein Kurztutorial: Verification of Neural Networks

Autorin: Stefanie Mühlberger

Datum 24.07.2020
Gut als erste Einführung geeignet.
Es gibt verständliche und motivierende Folien.

Kann als Thema nur vergeben werden, wenn ein weiterer Artikel dazukommt.

Liste der Titel

Hier nochmal in Kurzform als Liste nur mit den Titeln.

Thema 1:	DeepAbstract: Neural Network Abstraction for Accelerating Verification
Thema 2:	Maximum Resilience of Artificial Neural Networks
Thema 3:	Deep Save: A Data-driven Approach for Checking Adversarial Robustness in Neural Networks
Thema 4:	The Marabou Framework for Verification and Analysis of Deep Neural Networks
Thema 5:	Intriguing properties of neural networks
Thema 6:	Automated Verification of Neural Networks: Advances, Challenges and Perspectives
Thema XX:	– Nach Rücksprache