

Safety and Reliability for Learning Systems

Part V - Some latest/other approaches and outlook

Rüdiger Ehlers, Clausthal University of Technology

Marktoberdorf Summer School, August 2019

Some current directions of research

- Neural network verification
- Layer-by-layer propagations
- Symblicit testing
- Safe reinforcement learning
- Circumventing the lack of good specifications
- Learning easy to verify networks

Some current directions of research

- Neural network verification
- Layer-by-layer interpretation
- Symbolic testing
- Safe reinforcement learning
- Circumventing the lack of generalization
- Learning easy to verify networks

**Very Few
Examples Only!**

Neural network verification (1)

Wang et al., 2018

ReluVal:

- Symbolic range analysis for ReLU nodes
- Propagate lower and upper bounds (symbolic in the input variables to the network if possible)
- Split on ReLU phases to increase precision when needed

Neural network verification (2)

Katz et al., 2017, 2019

- Modified simplex algorithm (*first stage*) specifically for NN verification with ReLU non-linearities
- Introduces a special ReLU “fixing” step for candidate solutions in the simplex algorithm.
- Split on a ReLU phase if reasoning in circles.
- Katz et al. (2019) add:
 - Support for divide and conquer
 - Support for reasoning based on the network topology (e.g., symbolic bound tightening)

Combining formal methods & testing

Gopinath et al., 2018

- First, we can apply the network on training or random data
- Then, we can cluster the predictions for input similarity such that in each cluster, the predictions are all the same
- Clusters of the same prediction *may* represent regions of the input space in which the prediction is the same.
→ verify using a neural network verification approach

Layer-by-layer propagation methods

- Huang et al., 2017
- Xiang et al., 2017
- ...

Specification support

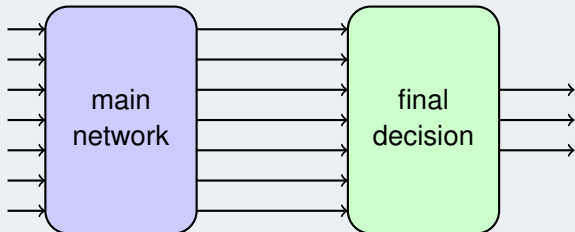
Cheng et al., 2019

- Idea: Learn a representation of the relevant cases along with the model
- Verify correct behavior for this representation model.

Specification support

Cheng et al., 2019

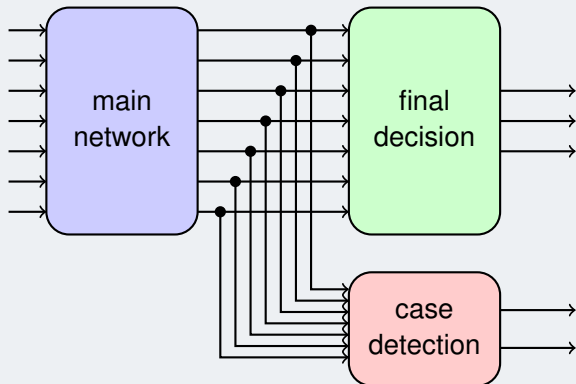
- Idea: Learn a representation of the relevant cases along with the model
- Verify correct behavior for this representation model.



Specification support

Cheng et al., 2019

- Idea: Learn a representation of the relevant cases along with the model
- Verify correct behavior for this representation model.



Learning easy to verify networks

Xiao et al., 2019

- Focus: Robustness checking of learned networks
- Use of special loss function that benefits networks for which ReLU phases change little around the training data points
- Reported speed-up of 4-13x in verification time

Making reinforcement learning safe

Other approaches

- Fulton and Platzer, 2018
- Phan et al., 2019
- ...



Outlook

Outlook

To achieve *scalability*

We need a co-development of **learning** and **verification** methods.

Outlook

To achieve *scalability*

We need a co-development of **learning** and **verification** methods.

To improve *correctness* of learned models

We need to find a way to integrate specifications into the learning process itself.

Outlook

To achieve *scalability*

We need a co-development of **learning** and **verification** methods.

To improve *correctness* of learned models

We need to find a way to integrate specifications into the learning process itself.

To achieve *impact*

We need to find ways to help with specifying the intended properties of learned models!

References I

- Chih-Hong Cheng, Chung-Hao Huang, Thomas Brunner, and Vahid Hashemi. Towards safety verification of direct perception neural networks. *CoRR*, abs/1904.04706, 2019. URL <http://arxiv.org/abs/1904.04706>.
- Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6485–6492. AAAI Press, 2018. URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17376>.
- Divya Gopinath, Guy Katz, Corina S. Pasareanu, and Clark W. Barrett. Deepsafe: A data-driven approach for assessing robustness of neural networks. In *Automated Technology for Verification and Analysis - 16th International Symposium, ATVA 2018, Los Angeles, CA, USA, October 7-10, 2018, Proceedings*, pages 3–19, 2018. doi: 10.1007/978-3-030-01090-4_1. URL https://doi.org/10.1007/978-3-030-01090-4_1.
- XiaoWei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. Safety verification of deep neural networks. In *Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I*, pages 3–29, 2017. doi: 10.1007/978-3-319-63387-9_1. URL https://doi.org/10.1007/978-3-319-63387-9_1.
- Guy Katz, Clark W. Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I*, pages 97–117, 2017. doi: 10.1007/978-3-319-63387-9_5. URL https://doi.org/10.1007/978-3-319-63387-9_5.
- Guy Katz, Derek A. Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljic, David L. Dill, Mykel J. Kochenderfer, and Clark W. Barrett. The marabou framework for verification and analysis of deep neural networks. In *Computer Aided Verification - 31st International Conference, CAV 2019, New York City, NY, USA, July 15-18, 2019, Proceedings, Part I*, pages 443–452, 2019. doi: 10.1007/978-3-030-25540-4_26. URL https://doi.org/10.1007/978-3-030-25540-4_26.
- Dung Phan, Nicola Paoletti, Radu Grosu, Nils Jansen, Scott A. Smolka, and Scott D. Stoller. Neural simplex architecture. *CoRR*, abs/1908.00528, 2019. URL <http://arxiv.org/abs/1908.00528>.

References II

- Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Formal security analysis of neural networks using symbolic intervals. In William Enck and Adrienne Porter Felt, editors, *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018.*, pages 1599–1614. USENIX Association, 2018. URL <https://www.usenix.org/conference/usenixsecurity18/presentation/wang-shiqi>.
- Weiming Xiang, Hoang-Dung Tran, and Taylor T. Johnson. Reachable set computation and safety verification for neural networks with relu activations. *CoRR*, abs/1712.08163, 2017. URL <http://arxiv.org/abs/1712.08163>.
- Kai Y. Xiao, Vincent Tjeng, Nur Muhammad (Mahi) Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing relu stability. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL <https://openreview.net/forum?id=BjfIVjAcKm>.