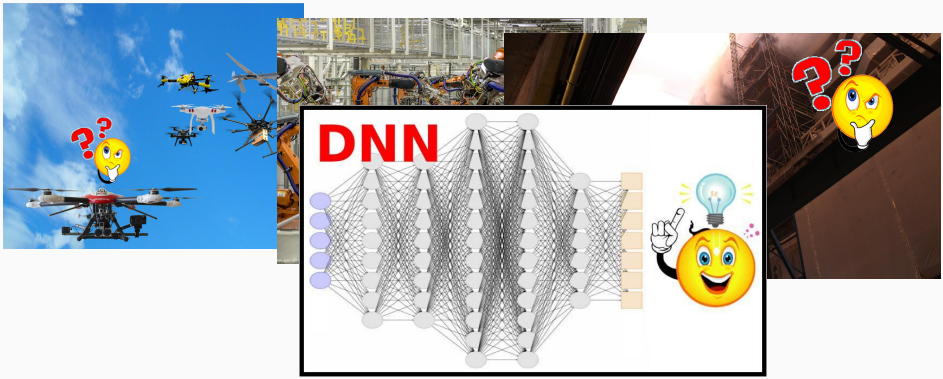
Neural Network Action Policy Verification via Predicate Abstraction

Marcel Vinzent, Jörg Hoffmann

SAARLAND UNIVERSITY
COMPUTER SCIENCE



e.g. PRL, e.g. [Toyer *et al.* (2018); Groshev *et al.* (2018); Issakkimuthu *et al.* (2018); Garg *et al.* (2019); Rivlin *et al.* (2020); Toyer *et al.* (2020)]

But what about trust in a learned neural network action policy?

- Explanation, e.g. [Chakraborti *et al.* (2019); Agogino *et al.* (2019)]
- Visualization, e.g. [Gros *et al.* (2020a)]
- Shielding, e.g. [Könighofer *et al.* (2017); Alshiekh *et al.* (2017); Fulton and Platzer (2018)]
- Testing, e.g. [Julian *et al.* (2020); Steinmetz *et al.* (2021)]
- Verification:
  - single decision episodes, e.g. [Katz *et al.* (2017); Gehr *et al.* (2018)]
  - n-step reachability, e.g. [Akintunde *et al.* (2018, 2019)]
  - statistical guarantees, e.g. [Gros *et al.* (2020b)]

Here: **safety verification**,

Do there exist $s \in S_0, t \in S_U$ such that t is reachable from s under an NN action policy π?

Predicate Abstraction

- State abstraction Θ^{α}:
  $\alpha\colon s \mapsto A$, while preserving transitions.
- Predicate abstraction $\Theta|_{\mathcal{P}}$ over predicates $\mathcal{P}$, e.g. [Graf and Saïdi (1997)],
  $s \mapsto A$ according to truth values s induces over $\mathcal{P}$,
  e.g. $\mathcal{P} = \{x = 7, x \leq y\}$.

Motivation: safety verification via (over-approximating) reachability analysis in $\Theta|_{\mathcal{P}}$.

Computing $\Theta|_{\mathcal{P}}$:

Does there exists a transition
from abstract state A to A' under action a in $\Theta|_{\mathcal{P}}$?

Do there exist concrete states $s \in A$ and $s' \in A'$ such that
there is a transition from s to s' under a in Θ?

– encoded as SMT-test [Barrett *et al.* (1994)].

Policy Predicate Abstraction

- $\Theta^{\pi}|_{\mathcal{P}}$, predicate abstraction of the *policy-restricted* state space Θ^{π}.

Does there exists a transition
from abstract state A to A' under action a in $\Theta^{\pi}|_{\mathcal{P}}$?

Do there exist concrete states $s \in A$ and $s' \in A'$ such that
there is a transition from s to s' under a in Θ and $\pi(s) = a$?

– SMT-test encodes NN-SAT problem.

Approach: Plugging in Progress on NN Analysis

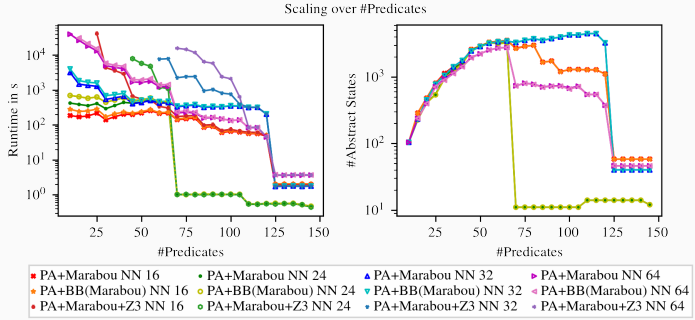
- over-approximating SMT-tests,
- dedicated NN analysis methods,
  e.g., *Marabou* [Katz *et al.* (2019)] for continuous over-approximation.

Experiments: Scalability Study

Experimental setting:

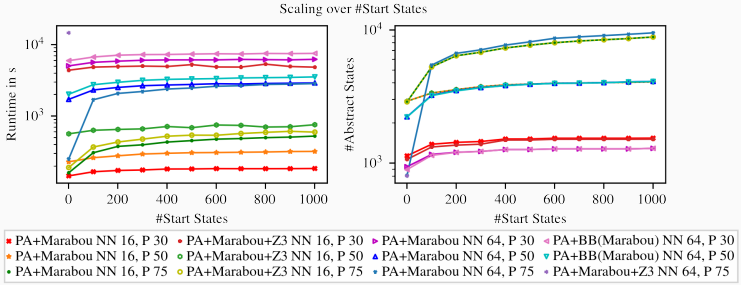
- *Z*3 [de Moura and Bjørner (2008)] & *Marabou* [Katz *et al.* (2019)]
- PA+Marabou (computing an over-approximation of $\Theta^{\pi}|_{\mathcal{P}}$),
- PA+Branch&Bound(*Marabou*) and PA+Marabou+Z3 (computing $\Theta^{\pi}|_{\mathcal{P}}$),
- Racetrack modeled in JANI [Budde *et al.* (2017)].

Scaling over $|\mathcal{P}|$



- PA+Marabou computes a rather fine over-approximation of $\Theta^{\pi}|_{\mathcal{P}}$.
- PA+Marabou/PA+BB(Marabou) outperforms PA+BB(Marabou)/PA+Marabou+Z3.
- Runtime is highly dependent on granularity of $\mathcal{P}$, increasing for coarse $\mathcal{P}$.

Scaling over $|S_0|$



Conclusion

- Policy predicate abstraction as a technique
  to enable NN action policy safety verification.
- Empirical results:
  - Safety verification via policy predicate abstraction may be feasible.
  - Rather fine but significantly less expensive over-approximations of $\Theta^{\pi}|_{\mathcal{P}}$ via PA+Marabou.

Questions?

References

Adrian Agogino, Ritchie Lee, and Dimitra Giannakopoulou. Challenges of explaining control. In *2nd ICAPS Workshop on Explainable Planning (XAIP'19)*, 2019.

Michael Akintunde, Alessio Lomuscio, Lalit Maganti, and Edoardo Pirovano. Reachability analysis for neural agent-environment systems. In *16th International Conference on Principles of Knowledge Representation and Reasoning (KR'18)*, pages 184–193, 2018.

M. E. Akintunde, A. Kevorchian, A. Lomuscio, and E. Pirovano. Verification of RNN-based neural agent-environment systems. In *AAAI19*, pages 6006–6013. AAAI Press, 2019.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. *CoRR*, abs/1708.08611, 2017.

References ii

C. W. Barrett, R. Sebastiani, S. A. Seshia, and C. Tinelli. Satisfiability modulo theories. *In Handbook of Satisfiability*, pages 825–885, 1994.

Carlos E. Budde, Christian Dehnert, Ernst Moritz Hahn, Arnd Hartmanns, Sebastian Junges, and Andrea Turrini. JANI: Quantitative model and tool interaction. In *TACAS (2)*, LNCS 10206, pages 151–168, 2017.

Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E. Smith, and Subbarao Kambhampati. Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In *Proceedings of the 29th International Conference on Automated Planning and Scheduling (ICAPS'19)*, pages 86–96. AAAI Press, 2019.

Leonardo de Moura and Nikolaj Bjørner. Z3: An Efficient SMT Solver. In C.R. Ramakrishnan and J. Rehof, editors, *Tools and Algorithms for the Construction and Analysis of Systems. TACAS 2008*, LNCS 4963, Berlin, Heidelberg, 2008. Springer. https://doi.org/10.1007/978-3-540-78800-3_24.

References iii

Nathan Fulton and Andreé Platzer. Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In *Proc. 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, 2018.

Sankalp Garg, Aniket Bajpai, and Mausam. Size independent neural transfer for RDDL planning. In *Proceedings of the 29th International Conference on Automated Planning and Scheduling (ICAPS 2019)*, pages 631–636, 2019.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin T. Vechev. AI2: Safety and robustness certification of neural networks with abstract interpretation. In *IEEE Symposium on Security and Privacy 2018*, pages 3–18, 2018.

S. Graf and H. Saïdi. Construction of abstract state graphs with PVS. In *9th International Conference on Computer Aided Verification (CAV)*, 1997.

References iv

Timo P. Gros, David Groß, Stefan Gumhold, Jörg Hoffmann, Michaela Klauck, and Marcel Steinmetz. TraceVis: Towards visualization for deep statistical model checking. In *Proceedings of the 9th International Symposium On Leveraging Applications of Formal Methods, Verification and Validation (ISoLA'20)*, 2020.

Timo P. Gros, Holger Hermanns, Jörg Hoffmann, Michaela Klauck, and Marcel Steinmetz. Deep Statistical Model Checking. In *Proceedings of the 40th International Conference on Formal Techniques for Distributed Objects, Components, and Systems (FORTE'20)*, 2020. available at https://doi.org/10.1007/978-3-030-50086-3_6.

Edward Groshev, Maxwell Goldstein, Aviv Tamar, Siddharth Srivastava, and Pieter Abbeel. Learning generalized reactive policies using deep neural networks. In *Proceedings of the 28th International Conference on Automated Planning and Scheduling (ICAPS'18)*, pages 408–416, 2018.

References v

Murugeswari Issakkimuthu, Alan Fern, and Prasad Tadepalli. Training deep reactive policies for probabilistic planning problems. In Mathijs de Weerdt, Sven Koenig, Gabriele Röger, and Matthijs T. J. Spaan, editors, *Proceedings of the 28th International Conference on Automated Planning and Scheduling (ICAPS)*, pages 422–430, 2018.

Kyle D. Julian, Ritchie Lee, and Mykel J. Kochenderfer. Validation of image-based neural network controllers through adaptive stress testing. In *23rd IEEE International Conference on Intelligent Transportation Systems (ITSC'20)*, pages 1–7, 2020.

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 *CAV (1)*, LNCS 10426, pages 97–117, 2017.

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 Kochenderfer, and Clark Barrett. The Marabou Framework for Verification and Analysis of Deep Neural Networks. In I. Dillig and S. Tasiran, editors, *Computer Aided Verification. CAV 2019*, LNCS 11561, Cham, 2019. Springer. https://doi.org/10.1007/978-3-030-25540-4_26.

Bettina Könighofer, Mohammed Alshiekh, Roderick Bloem, Laura Humphrey, Robert Könighofer, Ufuk Topcu, and Chao Wang. Shield synthesis. *Formal Methods in System Design*, 51(2):332–361, 2017.

Or Rivlin, Tamir Hazan, and Erez Karpas. Generalized planning with deep reinforcement learning. In *ICAPS 2020 Workshop on Bridging the Gap Between AI Planning and Reinforcement Learning (PRL)*, pages 16–24, 2020.

Marcel Steinmetz, Timo P. Gros, Philippe Heim, Danieal Höller, and Jörg Hoffmann. Debugging a policy: A framework for automatic action policy testing. In *Proceedings of the ICAPS Workshop on Bridging the Gap Between AI Planning and Reinforcement Learning (PRL'21)*, 2021.

Sam Toyer, Felipe Trevizan, Sylvie Thiebaux, and Lexing Xie. Action schema networks: Generalised policies with deep learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, 2018.

Sam Toyer, Sylvie Thiébaux, Felipe W. Trevizan, and Lexing Xie. Asnets: Deep learning for generalised planning. *Journal of Artificial Intelligence Research*, 68:1–68, 2020.