Stanford Center for AI Safety: Allstate Flagship Project #2

Project Title: Adaptive Stress Testing for Autonomous Vehicle Risk Assessment

1. Background

An approach known as adaptive stress testing (AST) has recently been used to find the most likely failures in aircraft collision avoidance systems [1], [2] and autonomous vehicles [3], [4]. AST was designed for safety-critical black-box systems, and thus requires minimal interaction with the system under test. Lee *et al.* [1] applied AST to aircraft collision avoidance systems using a modified Monte Carlo tree search algorithm to control the random number generator seed used by the system's simulation environment. By only controlling the seed of an otherwise intractable problem, AST found likely failure events in a system where failures are extremely rare. Koren *et al.* [3] applied AST to white-box autonomous vehicles using deep reinforcement learning as the solver and extended their approach to black-box autonomous vehicles using recurrent neural networks [4]. Koren and Kochenderfer [5] also propose a method to learn a heuristic reward using the go-explore algorithm, which otherwise can be difficult or infeasible to construct from domain knowledge.

To better understand failure events, approaches have been proposed to find human-interpretable failures [6] and to cluster high-level failures types into categories [7]. Corso and Kochenderfer [6] use a grammar based on signal temporal logic (STL), which is commonly used for falsification, to describe the failures and use genetic programming to optimize for high likely failure events. Their approach results in human-interpretable logical statements that describe the failures (e.g. "between 0 and 5 seconds the blinker is always on"). Lee *et al.* [7] propose a similar grammar-based approach for interpretability and also provide a categorization of failure events based on common behaviors described by the grammar. Their work aims to provide more useful descriptions of failure events to be addressed by the engineers and designers of the system under test. Other applications of AST include pairwise differential stress testing [2] and domain-relevant reward augmentation [8]. To stress test two different systems, Lee *et al.* [2] propose a pairwise approach for differential adaptive stress testing (DAST). Their approach uses reinforcement learning to encourage failures in one system and not the other to maximize the difference in their outcomes. DAST is useful when assessing a system relative to another baseline system and can also be used for regression testing. Corso *et al.* [8] encode domain relevant information into the AST reward function to encourage finding a larger and more expressive set of failures (to avoid exploiting repeatedly discovered failures).

One critical category of failure events is sensing and perception failures. An autonomous vehicle is equipped with an array of sensors for perception, such as cameras, GPS units, inertial navigation units, LiDARs, radars and transceivers for communication with other

vehicles on the road. To validate image-based neural network controllers, Julian, Lee, and Kochenderfer [9] use AST to apply disturbances to input images to find failures in an aircraft taxiing system. Their work addresses the validation of the multi-step properties in neural network controllers. Shetty and Gao [10] proposed an image-based cross-view geolocalization method for pose estimation with the aid of georeferenced satellite imagery. The approach consists of two Siamese neural networks that extract relevant features despite large differences in viewpoints, and integrates the crossview geolocalization output with visual odometry through a Kalman filter. In addition to images, there is also a large body of existing work by on LiDAR faults [11], [12], GPS faults [13], [14] and faults of sensor fusion [15].

2. Research Proposal

2.1 Overview

To perform autonomous vehicle risk assessment over black-box systems, we propose a framework for defining safety requirements through temporal logic specifications and using adaptive stress testing to validate each system under test. With this approach, we can define the failure events to search for (e.g., collisions) and the events that led to a failure (e.g., sensor failures) as human-interpretable temporal logic specifications. To better model real-world risk, our safety validation framework will account for realistic sensor uncertainties in the simulation, as well as interactions with other agents on the road (e.g., human-driven vehicles). Given the computational tractability of the risk assessment problem, we propose speed ups to adaptive stress testing by combining reinforcement learning techniques with local optimization methods.

2.2 Scope

The extent of this work is to build a cohesive risk assessment framework that is agnostic to the black-box autonomous vehicles under test. We will use off-the-shelf intelligent driver models [16] as our black-box autonomous vehicle systems and high-fidelity simulation environments like CARLA [17] for realistic integration of physical perception sensors. The design of the modular framework can easily include additional sensor types, agent behavioral models (e.g., Trajectron++ [18]), autonomous vehicle systems, and driving scenarios into the simulation.

2.3 Goals

Our objective is to provide a realistic framework for autonomous vehicle risk assessment. We aim to use adopted techniques from the safety validation literature, including adaptive stress testing, temporal logic specification robustness, agent behavioral models, and sensor uncertainty models. We will also design and develop algorithms to improve the computational tractability of the risk assessment.

3. Task Descriptions

Task 1: AST Modeling Framework

To expand on our existing research, we would like to incorporate STL robustness measurements into the AST reward function. For system requirements described using STL, robustness is a measure of how close the requirement is to being violated [19]. Current approaches in the falsification literature use robustness to guide their search [20]. To assess autonomous vehicle risk, system requirements can be specified in STL and we can use AST to find likely failures. These system requirements can be common across AV platforms, thus allowing us to appropriately compare any number of black-box systems. Combining the interpretability work described above with STL robustness, we could automate the process of finding measurable failures that are also human-interpretable.

Task 2: Uncertainty Models

We will aid AST with real-world perception as well agent behavioral uncertainties. Existing work such as [6] makes certain assumptions about the sensing noise profile (e.g. uniformly distributed in $[0, 1]$). We will extend the existing work with a real-world sensing profile, which includes not only nominal cases that form a multi-modal distribution, but also outlier characteristics. The more realistic noise profile will enable more accurate risk assessment, and thus higher level of safety. Another key source of uncertainty is represented by the presence of sentient agents (e.g., human-driven vehicles, pedestrian, etc.) on the road. We will again leverage STL, along with data-driven techniques (e.g., [18]), to develop high-quality simulation agents for the purposes of AST and risk assessments. This task will leverage our recent breakthroughs on imbuing logical structure into high-capacity, learning-based representations of human behaviors through a tool referred to as `stlcg` [21] .

Task 3: Computational Tractability

We would also like to propose to speed up the AST process. One proposed approach is extending AST to use a two-layered approach similar to Zhang *et al.* [22]. This approach splits the failure event search and the most likely failure event search into phases. The two-layered approach uses optimization techniques to search a local action-space and the result is fed to the higher-level reinforcement learner. Furthermore, we will apply prior knowledge about sensing and perception uncertainties in the autonomous vehicle's physical state to search for the most likely failure event first. For example, if an autonomous vehicle is driving west during sunset time, image failures due to sun glare may be the main failure cause and should be searched first. The goal is to help speed up falsification (i.e. finding failures) so the reinforcement learner can explore the failure-space for likely cases. Applying this technique to assess autonomous vehicle safety would ideally speed up the process and assess risk across a more diverse set of failures.

4. Timeline

	October	November	December
Task 1	Design and development of modular framework	Incorporate STL specifications	Incorporate robustness calculations
Task 2	Implement simple sensor models	Implement outlier detection	Devise STL-based simulation agents
Task 3	Design two-layer approach into framework	Implement local optimization method(s)	Test optimization methods
	January	**February**	**March**
Task 1	Incorporate STL specification interpretability	Test STL interpretability	Perform massive risk assessment tests
Task 2	Model GPS multipath effects and outlier measurement using a high-fidelity GPS simulator	Model vision-based perception and sensing with CARLA	Augment CARLA with realistic GPS sensing measurements and STL-based simulation agents
Task 3	Include IDM into framework	Interface with CARLA	Test end-to-end risk assessment
	April	**May**	**June**
Task 1	STL interpretability data collection	Analysis of STL failures	Formalize work into paper
Task 2	Expand to multi-modal sensing profiles including both vision and GPS	Port `stlcg` to Julia	Formalize work into paper
Task 3	Two-layered approach data collection	Analysis of RL/ optimization methods	Formalize work into paper

5. Deliverables

We anticipate the following deliverables:

1. A peer-reviewed conference paper per task submitted no later than July 15, 2021.
2. A final report on all tasks by July 15, 2021.
3. All general-purpose software will be made open source under a non-restrictive license (e.g., MIT or BSD) by July 15, 2021.

6. Researchers

1. Mykel Kochenderfer and Robert Moss, Stanford Intelligent Systems Laboratory
2. Grace Gao and Shubh Gupta, Stanford Navigation and Autonomous Vehicles Laboratory
3. Marco Pavone and Robert Dyro, Stanford Autonomous Systems Laboratory

References

[1] R. Lee, M. J. Kochenderfer, O. J. Mengshoel, G. P. Brat, and M. P. Owen, "Adaptive stress testing of airborne collision avoidance systems," in *Digital Avionics Systems Conference (DASC)*, 2015.

[2] R. Lee, O. J. Mengshoel, A. Saksena, R. Gardner, D. Genin, J. Silbermann, M. Owen, and M. J. Kochenderfer, *Adaptive stress testing: Finding likely failure events with reinforcement learning*, 2018. arXiv: `1811.02188`.

[3] M. Koren, S. Alsaif, R. Lee, and M. J. Kochenderfer, "Adaptive stress testing for autonomous vehicles," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018.

[4] M. Koren and M. J. Kochenderfer, "Efficient autonomy validation in simulation with adaptive stress testing," in *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2019.

[5] M. Koren and M. J. Kochenderfer, "Adaptive stress testing without domain heuristics using go-explore," *arXiv*, no. 2004.04292, 2020.

[6] A. Corso and M. J. Kochenderfer, "Interpretable safety validation for autonomous vehicles," *arXiv*, no. 2004.06805, 2020.

[7] R. Lee, M. J. Kochenderfer, O. J. Mengshoel, and J. Silbermann, "Interpretable categorization of heterogeneous time series data," in *Proceedings of the 2018 SIAM International Conference on Data Mining*, 2018.

[8] A. Corso, P. Du, K. Driggs-Campbell, and M. J. Kochenderfer, "Adaptive stress testing with reward augmentation for autonomous vehicle validation," in *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2019.

[9] K. D. Julian, R. Lee, and M. J. Kochenderfer, "Validation of image-based neural network controllers through adaptive stress testing," *arXiv*, no. 2003.02381, 2020.

[10] A. Shetty and G. X. Gao, "UAV pose estimation using cross-view geolocalization with satellite imagery," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[11] A. V. Kanhere and G. X. Gao, "LiDAR SLAM utilizing normal distribution transform and measurement consensus," in *Proceedings of ION GNSS+ Conference*, 2019.

[12] A. V. Kanhere and G. X. Gao, "Integrity for gps-lidar fusion utilizing a raim framework," in *Proceedings of ION GNSS+ Conference*, 2018.

[13] S. Gupta and G. X. Gao, "Particle RAIM for integrity monitoring," in *Proceedings of the 32nd International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2019), Miami, Florida*, 2019.

[14] S. Bhamidipati and G. X. Gao, "Multiple GPS fault detection and isolation using a graph-SLAM framework," in *31st International Technical Meeting of the Satellite Division of the Institute of Navigation, ION GNSS+*, 2018.

[15] S. Bhamidipati and G. X. Gao, "Integrity Monitoring of Graph-SLAM Using GPS and Fish-eye Camera," in *Journal of the Institute of Navigation*, 2020.

[16] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical Review E*, vol. 62, no. 2, Aug. 2000.

[17] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017.

[18] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *European Conf. on Computer Vision*, In Press, 2020.

[19] G. E. Fainekos and G. J. Pappas, "Robustness of temporal logic specifications for continuous-time signals," *Theoretical Computer Science*, vol. 410, no. 42, 2009.

[20] A. Corso, R. J. Moss, M. Koren, R. Lee, and M. J. Kochenderfer, *A survey of algorithms for black-box safety validation*, 2020. arXiv: 2005.02979.

[21] K. Leung, N. Aréchiga, and M. Pavone, "Back-propagation through STL specifications: Infusing logical structure into gradient-based methods," in *Workshop on Algorithmic Foundations of Robotics*, 2020.

[22] Z. Zhang, G. Ernst, S. Sedwards, P. Arcaini, and I. Hasuo, "Two-layered falsification of hybrid systems guided by Monte Carlo tree search," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, 2018.