Towards Scalable Coverage-Based Testing of Autonomous Vehicles

James Tu, Simon Suo, Chris Zhang, Kelvin Wong, Raquel Urtasun

AV Testing
- Autonomous Vehicles are often tested in simulation through parameterized scenarios
- Each parameter combination yields a concrete scenario
- Pass/fail from thresholding underlying continuous metric

Our Testing Framework - GUARD
- Gaussian Process (GP) leverages observed concrete scenarios tests to estimate the probability of passing across the parameter space
- Use a probability threshold to partition the space into pass/fail/unknown
- Coverage = percent of parameter space that is not unknown
- Samples near the pass/fail boundary are more informative
- Samples where the GP is uncertain is more informative
- Testing process - iteratively sample concrete scenarios using these two criteria, update GP model

Comparison With Baselines
- Coverage: percent of parameter space that is not unknown
- Balanced accuracy: how accurate pass/fail predictions are, correct class imbalance since fails are much more rare
- Error recall: percent of ground truth failures that are predicted to be fail by the GP, useful for autonomy development
- False positive rate: percent of predicted passes that are correct. Incorrectly predicting passes can be detrimental to safety

Task
- Goal of testing - understand if the AV will pass or fail on concrete scenarios across parameter space
- Difficult to directly cover the continuous space, because infinitely many concrete scenarios
- Need to leverage observed test outcomes to estimate the outcome on unseen tests
- Task: execute a finite set of concrete scenarios and partition the parameter space into 3 regions: pass, fail, unknown

GUARD In Practice
- GUARD is able to benchmark two versions of the AV and compare their safety performance
- Can discover scenarios where the system regressed

Qualitative Results
- 2D slice of 5D parameter space
- Existing methods limited by discretization of parameter space