Predicting Safety-Critical Misbehaviours in Autonomous Driving Systems using Autoencoders

Università della Svizzera italiana

Software Institute

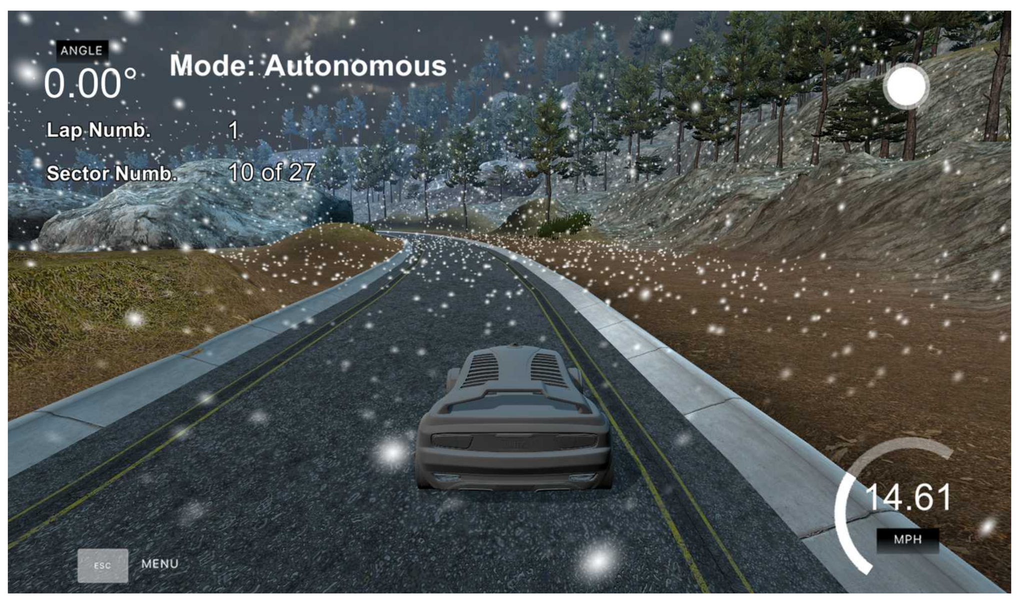
Andrea Stocco, Michael Weiss, Marco Calzana, Paolo Tonella

Motivation

We address the problem of recognizing unexpected execution contexts with the purpose of predicting potential safety-critical misbehaviours. Our approach SelfOracle is based on a novel concept of self-assessment oracle, which monitors the DNN confidence at runtime, to predict unsupported driving scenarios in advance. Our approach is based on the reconstruction error of autoencoders as a black-box confidence metric.

Improved Simulator

We implemented two additional components within the Udacity simulator.
The **unexpected context generator** gradually injects unseen conditions in autonomous driving mode (i.e., conditions diverse from the training mode's defaults). Instances of these situations deal with *illumination* (day/night cycle) or *weather* (rain, snow, fog), as well as their possible combinations.



The **collision/OBE detection system** records any unwanted interaction of the self-driving car with the environment during testing (e.g., collisions, or car driving off track). The result is a set of labeled images that we can use to experiment the effectiveness of SelfOracle at anticipating such unexpected scenarios.

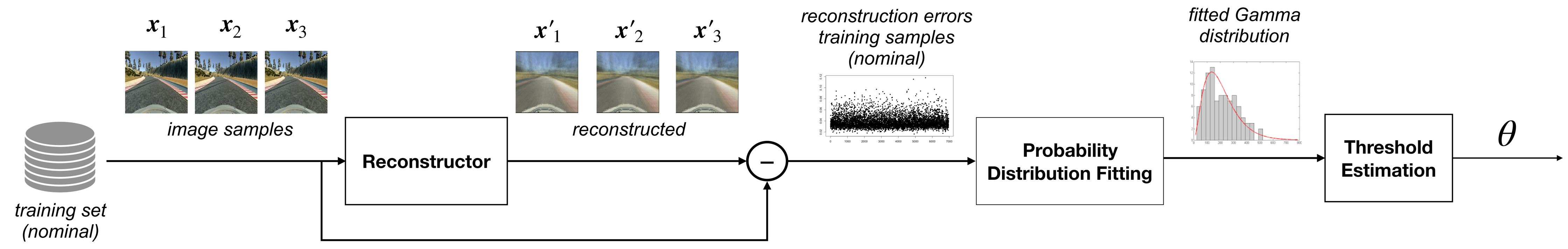
Links

SelfOracle is available at
`https://github.com/testingautomated-usi/selforacle`

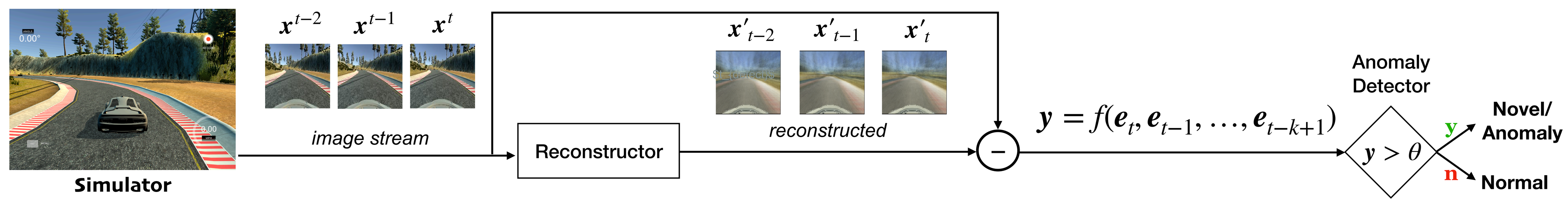
Our improved Udacity simulator is available at
`https://github.com/tsigalko18/self-driving-car-sim/releases/tag/USI_v1.0.0`

SelfOracle

Our approach combines a reconstruction-based anomaly detector with a predictive model of normality based on time series analysis. In particular, we used as reconstructors four autoencoders: (1) *SAE* (simple autoencoder with a single hidden layer), (2) *DAE* (deep five layers fully-connected autoencoder), (3) *CAE* (convolutional autoencoder alternating convolutional and max-pooling layers), and (4) *VAE* (variational autoencoder).



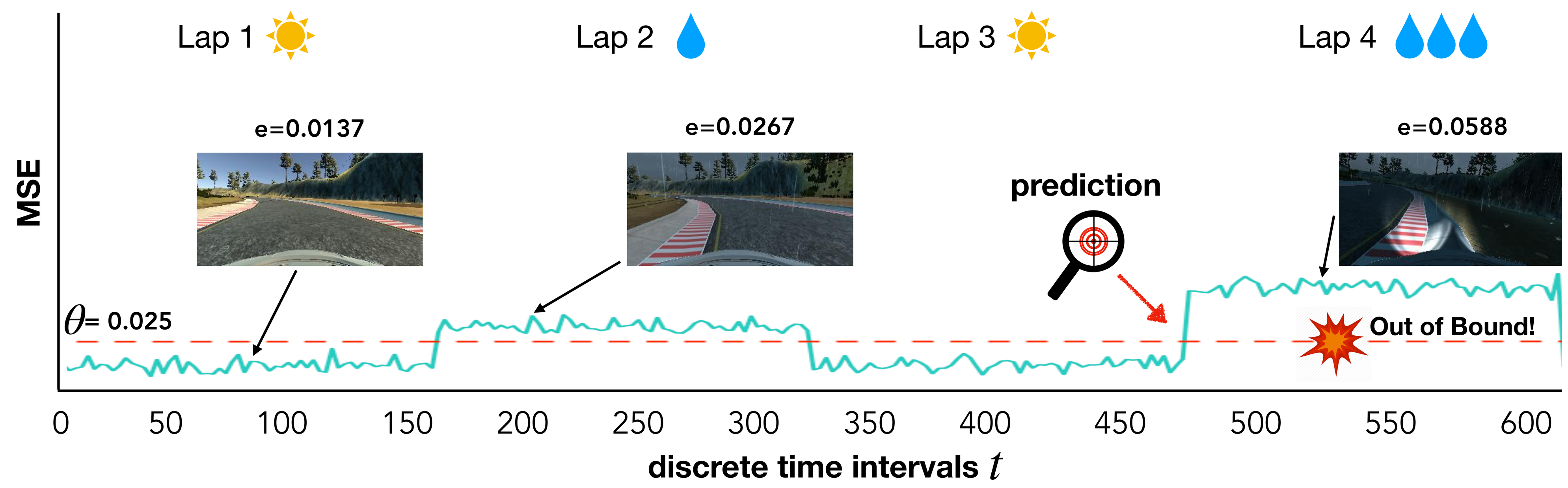
First, we record data representative of the behaviour of the car when driving in nominal conditions. Second, we fit a probability distribution of the nominal behaviours and set a confidence level (i.e., a threshold) that defines the acceptable false positive rate.



We inject anomalous/unseen conditions in the simulator, and we re-execute the self-driving car, recording each failure of the self-driving component, namely out-of-bounds and collisions. Then, we evaluate whether our predictive model is able to signal the occurrence of such failures in advance.

Evaluation

Our evaluation data consist of 72 simulations. We ran all tested self-driving car models on all available tracks under all conditions. Overall, we obtained a dataset of 778,592 images. We split the evaluation set into *windows of consecutive frames*, which we labelled as either anomalous or normal. The goal of SelfOracle is maximizing the prediction of shortly-following misbehaviors in anomalous windows (true positives), while minimizing the false alarms, i.e., wrong misbehavior predictions in normal windows (false positives).



Effectiveness. In our experiments, the best reconstructors are VAE and SAE, with comparable overall performance. At $\epsilon = 0.05$, VAE predicts correctly 589/765 misbehaviours (77%), with 84/765 false alarms (11%) due to adverse conditions that were not that extreme to make the system fail. This was expected, since we are measuring FPR in tracks with injected anomalies. DeepRoad$_{IV}$ predicts correctly 252/765 misbehaviours (33%), with 76/765 false alarms (10%). Thus, VAE detected 337 more misbehaviours, with a comparable false alarm rate. At $\epsilon = 0.01$, SAE predicts correctly 451/765 misbehaviours (59%), with 38/765 false alarms (5%). DeepRoad$_{IV}$ predicts correctly 153/765 misbehaviours (20%), with 46/765 false alarms (6%). Thus, SAE detected 298 more misbehaviours, again with a comparable false alarm rate.

Prediction. In our experiments, all configurations of SelfOracle are able to predict, on average, an upcoming misbehaviour up to 60 frames (around 6 s) in advance.

Comparison. To summarize, in our experiments SelfOracle has shown to be more effective than DeepRoad$_{IV}$ at predicting misbehaviours. Results of AUC-PRC and AUC-ROC show significant improvements across all thresholds, regardless of the technique being used and the reaction period considered. Concerning the performance, in our experiments, the autoencoders took ≈3 ms per prediction whereas DeepRoad took ≈45 ms per prediction (+1400% increment). While both runtime measures may seem acceptable in practical scenarios, it is worth remembering that DeepRoad$_{IV}$ requires to dramatically sub-sample the training set available for the experiments to achieve such execution times. Indeed, only few hundreds images can be used, because the technique behind DeepRoad$_{IV}$ is computationally very expensive. Hence, differently from our approach, it is also quite unlikely to scale to training datasets used by industry manufacturers.