Towards Safe Reinforcement Learning via OOD Dynamics Detection in Autonomous Driving System (Student Abstract)

Arnaud Gardille[1], **Ola Ahmad**[2]

[1]Paris-Saclay University, France
[2]Thales Digital Solutions, Montreal, Canada
arnaud.gardille@universite-paris-saclay.fr ola.ahmad@thalesdigital.io

Abstract

Deep reinforcement learning (DRL) has proven effective in training agents to achieve goals in complex environments. However, a trained RL agent may exhibit, during deployment, unexpected behavior when faced with a situation where its state transitions differ even slightly from the training environment. Such a situation can arise for a variety of reasons. Rapid and accurate detection of anomalous behavior appears to be a prerequisite for using DRL in safety-critical systems, such as autonomous driving. We propose a novel OOD detection algorithm based on modeling the transition function of the training environment. Our method captures the bias of model behavior when encountering subtle changes of dynamics while maintaining a low false positive rate. Preliminary evaluations on the realistic simulator CARLA corroborate the relevance of our proposed method.

Introduction

Deep reinforcement learning has achieved superhuman performance in most Atari games, so why is this promising approach not used so much in the industry? An end-to-end autonomous driving system trained with such a method could become much cheaper and more efficient than the contemporary highly modular hand-designed approach. In addition to proving its feasibility, the method proposed in (Chen, Xu, and Tomizuka 2020) offers a human-understandable interpretation of the agent's decisions.

However, like other machine learning methods, reinforcement learning is based on statistics. Therefore, it is susceptible to unexpected behaviors when it faces transitions out of its training distribution (OOD). Although several methods can help to deal with this crucial problem (Zhao, Queralta, and Westerlund 2020), an effective method to quickly detect when the environment's transitions differ from those of agent training appears to be a prerequisite for the use of DRL in safety-critical systems. Our preliminary experiments on the CARLA simulator show how a minor modification to the car's engine or a change in the weather could cause a trained RL agent to make harmful decisions. Several methods considered the OOD detection problem on raw inputs such as images, but to our knowledge, no method has been developed explicitly for detecting drifts in the environment's dynamic. Deep learning methods can estimate the agent network's confidence, which can be used as an OOD metrics. However, they require a particular training procedure and are not sufficiently robust to detect anomalous environment's dynamic because of spurious correlations in data (Mohammed and Valdenegro-Toro 2021).

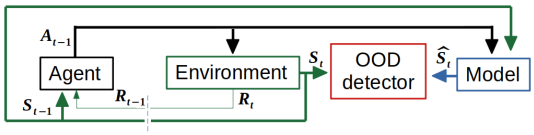

Figure 1: Method's architecture

Our proposed method is illustrated in Figure 1. We propose to model the dynamics of the current environment and compute an OOD score using the model's outputs and the current state. We then use a statistical threshold to capture anomalous changes (or drifts) in the dynamics distribution.

Method

We assume that the training environment's dynamics can be modeled by a transition function $\mathcal{T}$ such as $\mathcal{T}(S_{t-1}, A_{t-1}) = S_t$. We create several in-distribution trajectory datasets $((S_{t-1}, A_{t-1}), S_t)$. The first one is used to train a supervised model M to mimic the transition function: $M(S_{t-1}, A_{t-1}) \approx S_t$. The prediction error over the training environment represents a system's noise, which can be modelled as a zero-centred gaussian distribution:

$$D_t := M(S_{t-1}, A_{t-1}) - S_t \sim \mathcal{N}(0, \sigma_t^2) \tag{1}$$

where σ_t being a function of (S_t, A_t). In an OOD environment, slightly different dynamics will reasonably cause D_t to be biased. To measure this bias, we deploy the agent on the targeted environment, and calculate D_t at each step along a predefined interval $[T-t, ..., t]$. Once we obtain sufficient number of samples T, we verify the assumption in eq. (1). If σ_t is stationary (i.e., $\sigma_t = \sigma$ independent from t), we could simply use a student's t-test (τ) to reject the hypothesis that the prediction error is drawn from a normal distribution

$$\tau = \frac{\bar{D}_t}{\widehat{\sigma}/\sqrt{T}} \tag{2}$$

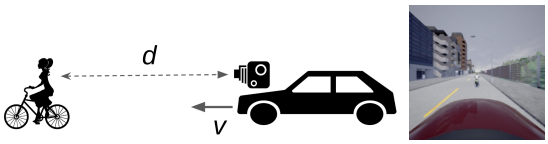

Figure 2: Context of the experimentation

where $\bar{D}_t$ is the samples mean and $\widehat{\sigma}$ the standard deviation over the samples $(D_t)_{T-t,\dots,t}$. However, the RL environment exhibits complex uncertainty causing non-stationarity in model error characteristics. To tackle this issue, after training M, we also train a regression model $g(S_{t-1}, A_{t-1}) \approx \sigma_t$ to predict the variance on another in-distribution dataset. Then, we replace $\widehat{\sigma}$ by $g(S_{t-1}, A_{t-1})$ in (2). In an OOD scenario, the models g and M are likely to make significantly incorrect predictions which will then be detected by the statistical test.

Experimental Results

We evaluate our method on automatic emergency breaking scenario (Figure 2). It is developed on a gym embedding of CARLA, a realistic autonomous driving simulator. At each time step of 0.1s, the agent perceives its speed v_t and the distance d_t to the frontal obstacle, then decides to brake or accelerate. Using RGB images, we train a ResNet18 model to estimate the distance d_t. We configured the training environment to include clear midday weather and allow for OOD cases during testing. Due to the latency between the action of accelerating and the actual acceleration provided by CARLA, the state vector must include a history of previous distances, speeds, and actions to ensure an optimal transition function. The reward increases when the agent's speed becomes stable around a target value and becomes prohibitively low when the agent hits the obstacle. We trained a Soft Actor Critic agent to behave efficiently in this scenario. Our models of the dynamics and variance function are fully connected MLP neural networks, trained with mean square loss functions for regression. In our experiments, we intervened in the training environment to sufficiently disrupt the trained agent such that it sometimes collides with the obstacle. The first intervention implies increasing the car's acceleration by 10%, which is hard enough to be detected as an OOD but sufficient to confuse the RL agent. In the second experiment, we changed the weather to a hard rain sunset, which caused the feature extractor to overestimate the distance d_t. As shown in Figure 3, our method efficiently detects both interventions as OOD events. Moreover, we obtained an AUC score of 0.82 and 0.93 respectively on these experiments.

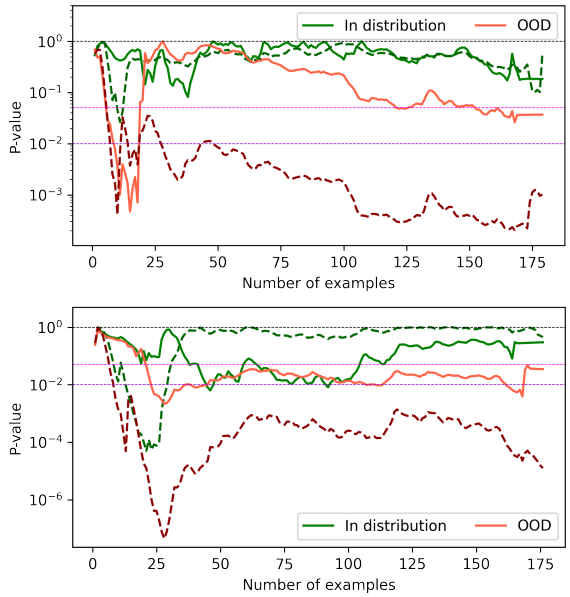

Figure 3: Evolution of the p-values of the tests depending on the sample's size T for the first experiment on top, and second experiment on the bottom. The test with and without normalizing by the approximated standard deviations are respectively in dashed and continuous lines. The pink and violet lines respectively correspond to a p-values of 5% and 1%. We can consider the tests to be significant when $T > 50$.

Conclusion and Perspectives

We argue that end-to-end monitoring of DRL agents in autonomous driving systems is necessary to ensure safety. This work focused on out-of-distribution detection to analyze the agent's bad behavior under the emergency breaking scenario. To show that our method is efficient and reliable, we introduced two OOD tasks: one concerning the internal dynamics of the car, and the second is a malfunction caused by perceptual OOD events. Furthermore, we show that normalizing the prediction error of the dynamics with an estimate of their variance improves the method's sensitivity to subtle environmental changes. Although our OOD detection method relies on modeling the transition function of the environment, it can be generalized to process the image inputs directly. For example, instead of using a perception model to extract meaningful low-dimensional semantics, a generative model, like the one used in (Chen, Xu, and Tomizuka 2020), can be used to predict the next image, then use our method to define the OOD score. For future work, we plan to evaluate our method on the benchmark proposed in (Mohammed and Valdenegro-Toro 2021) and investigate a runtime monitoring framework for OOD detection in DRL.

References

Chen, J.; Xu, Z.; and Tomizuka, M. 2020. End-to-end Autonomous Driving Perception with Sequential Latent Representation Learning. *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Mohammed, A. P.; and Valdenegro-Toro, M. 2021. Benchmark for Out-of-Distribution Detection in Deep Reinforcement Learning. In *Deep RL Workshop NeurIPS 2021*.

Zhao, W.; Queralta, J. P.; and Westerlund, T. 2020. Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 737–744.