Ensemble based uncertainty estimation with overlapping alternative predictions

Anonymous Author(s)
Affiliation
Address
email

Abstract

A reinforcement learning model will predict an action in whatever state it is - even if there is no distinct outcome due to unseen states the model may not indicate that. Different methods for uncertainty estimation can be used to indicate this. Although, uncertainty estimation is a well understood approach in AI, the overlap of effects like alternative possible predictions (multiple feasible actions in a given state) in reinforcement learning is not so clear and to our knowledge, not so well documented in current literature. In this work we investigate uncertainty estimation on simplified scenarios in a gridworld environment. Using model ensemble based uncertainty estimation we propose an algorithm based on action count variance to deal with discrete action spaces and delta to ID action variance calculation to handle overlapping alternative predictions. To visualize the expressiveness we create heatmaps for different ID and ODD scenarios on gridworlds and propose an indicator for uncertainty. We can show, that the method will indicate potentially unsafe states when the agent is near unseen elements in the scenarios (OOD) and can distinguish between OOD and overlapping alternative predictions.

1 Introduction

In order to apply Machine Learning in today's and future real world applications it is necessary to take into account the given system boundaries of an applications, e.g. resource and/or timing requirements and safety constraints.

There is a high incentive to use Reinforcement Learning (RL) in parts of these system approaches, as to the high flexibility and self-handling in decision making tasks. However, this also comes with a high risk in safety critical applications. As an RL agent will predict an action in whatever state it will find itself, it is of high importance for the application controller to be aware of the certainty and confident of its own decisions in all situations it could face. This includes situation form inside as form outside the training data.

To overcome this problem in the larger field of safe RL this paper focuses on estimating uncertainty from data during deployment (test data) that is out of the distribution of the training data (distributional shift) as one aspect of the problem solution.

Distributional shift in data science is widely understood as the distributional difference between training and test data of an algorithm. This distributional shift originates from different sources, such as perturbations to the data-set due to system complexity and a lack of training data. In machine learning, a distributional shift often leads to degraded performance during test, because the probability distribution over data (or state-action pairs in RL) is shifted and a classifier or an agent may therefore predict wrong or sub-optimal classes (or actions). In general, when the testing distribution differs

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

from the training distribution, machine learning systems may not only demonstrate poor performance, but also wrongly assume that their performance is good.

This even more applies to system solutions based on Reinforcement Learning. To overcome this limitation, safe Reinforcement Learning (Safe RL) solutions must be capable of detecting and handling the uncertainty in the decision-making process. Therefore, it is necessary to deal with uncertainty and potentially improve given algorithms to optimize these in respect to safety constraints. For instance, uncertainty estimation can detect a lack of generalization due to insufficient training and unseen states during training (OOD, epistemic uncertainty) as well as uncertainty resulting from randomness in the environment (aleatoric uncertainty).

For epistemic uncertainty, a set of alternatively trained agents can be used to predict on states. In states, where the agents are unsure due to a lack of training, the different agents will randomly predict different actions, due to a lack of substance of the prediction. This can be utilized to indicate uncertainty, but possible alternative predictions may overlap the deviation of the predictions. Therefore, it is necessary to differentiate between the two effects.

2 Related work

There is a lot of work in recent years in respect to OOD in the classical image classification domain as well as some work in the RL domain.

Pimentel *et al.* [2014] define novelty as test data, that differ in some respect from data during training. A model with a lack of training data is less capable in respect to generalization to conceptual not new but different looking data during test. As similar, Hendrycks and Gimpel [2018] define OOD as test data, which is from a different distribution as the training data and describe OOD detection as a threshold –based process. There is a differentiation between OOD as to be near and far from the training data distribution. The closer OOD is to the training data distribution the more it comes to a lack of training data only, but the further away it gets it rather represents conceptual or semantic differences, like samples which are completely out of the given classifications. In this regard, DNNs tend to be overconfident in predictions on unseen data and can give unpredictable results for far-from-distribution test data Lütjens *et al.* [2019].

A lot of prior work focuses on OOD as a concept of samples from outside the class set. If samples are outside the set, classifications for that cannot be learned, even with more/unlimited training. So one often seen approach is, to specify a separate OOD class to train the model on Pimentel *et al.* [2014]. Similarly, DeVries and Taylor [2018], Mohseni *et al.* [2020] define OOD samples as examples for classes different from those in the ID dataset. Lee *et al.* [2018] describe ID as a trained distribution by a classifier and OOD as sufficiently different from it. Also, Schwaiger *et al.* [2020] follow the approach to consider a strong difference between training and test data to be OOD. They describe ID data as conceptually similar to data trained on and OOD data as differ strongly from training data. Yu and Aizawa [2019] go as far as to define OOD by the gap in between classified ID data. They propose to maximize the discrepancy between the decision boundaries of e.g. two classifiers to push OOD samples outside. They also follow the concept of near and are far from the distribution.

Furthermore, it is important to understand the difference of epistemic and aleatoric uncertainty for uncertainty estimation as a proxy for OOD detectors. Epistemic uncertainty arises out of a lack of sufficient data to exactly infer the underlying system Sedlmeier *et al.* [2019]. Clements *et al.* [2020] claim epistemic uncertainty to stem from limited data. Epistemic uncertainty can indicate samples that reside far away from the data distribution as well as the capabilities of a model generalizes to data close to it Postels *et al.* [2020]. In contrast, aleatoric uncertainty arises from stochastic environments and must be accounted for in risk-sensitive applications Clements *et al.* [2020], Chua *et al.* [2018]. Aleatoric uncertainty cannot be solved just by more training. For safety reasons, aleatoric uncertainty is a significant indicator in RL and should also be considered. However, in this paper we focus on the detection of epistemic uncertainty, as we focus on distributional shift and OOD detection.

Wang *et al.* [2019] investigate a distributional RL algorithm D3PG, which models the uncertainty in the form of a return distribution in which the expected value is the Q-value. Different actions might be used when the distribution is bimodal or multimodal depending on the application scenario.

Kahn *et al.* [2017] present an uncertainty-aware model-based learning algorithm that estimates the probability of collision together with a statistical estimate of uncertainty. The predictive model is

based on bootstrapped neural networks using dropout. In regions of high uncertainty, their risk-averse cost function causes the robot to revert to a cautious low-speed strategy. In Silva *et al.* [2020] they propose an action-advising framework where the agent asks for advice when its epistemic uncertainty is high for a certain state to accelerate reinforcement learning. They add as a last layer multiple heads estimating separately expected values for each action, as done in Bootstrapped DQN. As the learning algorithm updates the network, their predictions get closer to the real function, and one close to the others. Sedlmeier *et al.* [2019] use uncertainty based OOD, using Q-value uncertainty in DQN Algorithm. They compare MC-Dropout, Bootstrapped and Bootstrapped with prior functions. They also address the problem of alternative valid predictions of the model agent. Unfortunately, they do not dig into detail, when it comes to the uncertainty estimation in those cases but rather calculate an overall estimate for the epoch.

Hoel *et al.* [2020] estimate uncertainty for RL based on ensembles with randomized prior functions (RPF). They are based on Osband *et al.* [2018] and propose a criterion function. They choose safe actions in unknown situations far from the training distribution. In Hoel *et al.* [2021] they also utilize an ensemble of DQN agents to estimate Q-value uncertainty to switch back to a fallback policy in uncertain situations given a certain threshold. An ensemble is trained on bootstrapped data, which provides a distribution over the estimated Q-values to provide a Bayesian estimation of the epistemic uncertainty. The epistemic uncertainty estimate is then be used to choose less risky actions in unknown situations. However, they don't take into account a potential overlapping uncertainty due to possible alternative actions.

3 Background

3.1 Reinforcement Learning and MDP

In Reinforcement Learning, we consider an agent that sequentially interacts with an environment modeled as a Markov Decision Process (MDP). An MDP is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \mu_0)$ composed by the set of states $\mathcal{S}$, the set of actions $\mathcal{A}$, the reward function $\mathcal{R}$, the transition probability function $\mathcal{P}$ which describes the system dynamics, and the μ_0 is the starting state distribution. The transition probability function $\mathcal{P}(s_{t+1}|s_t, a_t)$ models the probability of transitioning to a state s_{t+1} given a previous state s_t and taking the action a_t.

The return as sum of the discounted reward is given by

$$\mathcal{R}_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}. \tag{1}$$

In the MDP framework, at each timestep, the agent observes the current state, takes an action, transitions to the next state drawn from the distribution, and receives a reward. The action-value function, also known as the Q-value function, represents the expected return when following a policy π (which basically maps states into actions, i.e., $\mu : S \mapsto A$), as shown below.

$$\mathrm{Q}^{\pi}(s, a) = \mathbb{E}[R_t | s_t = s, a_t = a, \pi]. \tag{2}$$

Q-learning, in which a policy is learned using Q-values, is a popular model-free method. Deep Q-networks (DQNs) are an extension of Q-learning that work with neural networks and optimization algorithms commonly used in deep learning, such as gradient descent. To do so, the temporal-different error δ_t can be derived from the Q-value function using the Bellman operator, resulting in the equation below, whereas θ^- and θ are the DQN parameters from the target and the prediction network, respectively.

$$\delta_t = r_t + \gamma \max_a Q\left(s_{t+1}, a; \theta^-\right) - Q\left(s_t, a_t; \theta\right). \tag{3}$$

3.2 Distributional shift and OOD

Distributional shift and OOD are two concepts that are closely related. For instance, Hendrycks and Gimpel [2018] correlate distributional shift to OOD by describing the latter as samples drawn from a

different distribution than the training data. When transferring this definition to RL problems, it is important to distinguish distributions that are closer or further away from the training distribution. It is expected that an RL agent would be able to perform well in scenarios that are slightly different from the training ones, i.e., it should be able to generalize, but when the situation is too dissimilar (perhaps at a semantic level) the agent might have its ability to make proper decisions severely affected. A model or agent which is not able to detect a distributional shift may tend to be overconfident in predictions on unseen data and can give unpredictable results for far-from-distribution test data Lütjens *et al.* [2019].

Some authors also show that epistemic uncertainty can be used as a proxy for detecting distributional shifts, e.g., Sedlmeier *et al.* [2019]. Epistemic uncertainty is usually associated with a lack of sufficient data to better infer the underlying system. Clements *et al.* [2020] claim that epistemic uncertainty stems from limited data. However, besides detecting a models lack of generalization to data close to it epistemic uncertainty can also point to samples that reside far away from the data distribution known to the trained model, as shown in Postels *et al.* [2020].

Defining distributional shift within the RL domain is not trivial Haider *et al.* [2021]. In this paper we assume that distributional shift can be characterized by changes in the system dynamics, more specifically the shift of the distribution over the state transitions given state action pairs between training and test in MDPs, as shown below:

$$\mathrm{P}_{train}[s_t + 1|s_t, a_t] \neq P_{test}[s_t + 1|s_t, a_t]. \tag{4}$$

Additionally, when considering partially observable MDPs (POMDPs) where the system's state cannot be assessed but rather an observation o_t is available to the agent, the shift of the distribution over observations given states has to be taken into account:

$$\mathrm{P}_{train}[o_t|s_t] \neq P_{test}[o_t|s_t]. \tag{5}$$

3.3 Ensemble based uncertainty estimation

As discussed previously, uncertainty estimation often is the utilized method to detect a distributional shift. Detecting a distributional shift for the sake of exploration during training overlaps with the concept of novelty detection. In contrast, applying distributional shift detection during test time is considered as OOD detection in the classical image classification domain, where OOD describes the distributional shift of the data distribution between training and test.

This paper focuses on ensemble based epistemic uncertainty estimation to detect distributional shift and OOD data during test time. An ensemble of trained agents on a subset of the available data will estimate with little variance in well trained states. When the ensemble members face little trained states, the estimates vary naturally across the members and give a distribution over the estimated Q-values. The variance of the estimated Q-values can be used to quantify the epistemic uncertainty of a decision.

An ensemble on bootstrapped data over DQNs provides a distribution over the estimated Q-values to provide a Bayesian estimation of the epistemic uncertainty. The Q-values will converge to the real values in situations the agent learned sufficient. In untrained situations, the Q-value estimates will still diverge and the variance will therefore give an estimate of the epistemic uncertainty.

In Osband *et al.* [2018] they use random prior functions to introduce diversity in an ensemble of agents trained on bootstrapped data. The expected return is than given by

$$Q_k(s, a) = f\left(s, a; \theta_k\right) + \beta p(s, a; \hat{\theta}_k) \tag{6}$$

In Hoel *et al.* [2020] and Hoel *et al.* [2021] they use the variance of the Q-values of the ensemble estimate to derive an uncertainty estimation threshold to invoke a backup policy. With the variance $Var_k[Q_k(s, a)] < \sigma^2$ the policy with threshold can be calculated by

$$\pi_\sigma(s) = \begin{cases} \arg\max_a \mathbb{E}_k[Q_k(s, a)] & \text{if } Var_k[Q_k(s, a)] < \sigma^2 \\ \pi_{backup}(s) & otherwise \end{cases} \tag{7}$$

4 Ensemble Uncertainty estimation based on action count variance and delta to ID

4.1 Action count variance uncertainty estimation

The Q-value is a continuous variable where high variance in the predictions means high uncertainty of the ensemble. However, when given encapsulated agents or due to other reasons, where the Q-values are not accessible, it is possible to use the action prediction result from the ensemble members and derive an uncertainty estimate. Instead, it is possible to take the deviation on proposed actions of an ensemble, to indicate uncertainty. In cases where the action space is continuous, the variance can be directly calculated as action variance like with the Q-values. However, with discrete action spaces, this will lead to false results, as the actions themselves are orthogonal and e.g. a mean action can not be calculated. Therefore, in cases where the action space is discrete, we propose to calculate an action count on each action over the ensemble given a certain state and then calculate the variance of that action count (ACV - action count variance). When the ACV is low, there is a balance in the proposed different actions over the ensemble and the uncertainty is therefore high. In contrast, when the action count variance is high, there is a concentration of one or more actions in the ensemble and the uncertainty is low. The higher the ACV gets, the lower the uncertainty. Given $N_{actions}$ as the number of alternative actions and $K_{ensemble}$ as the number of ensemble members the minimum of the ACV gets to equation 8 and the maximum to equation 9.

$$Var_{min}[AC(s,a)] = \frac{N_{actions}^2}{K_{ensemble}} \tag{8}$$

$$Var_{max}[AC(s,a)] = \frac{N_{actions}}{K_{ensemble}} \tag{9}$$

A backup policy can then be chosen based on the ACV calculation like given in equation 10.

$$\pi_\sigma(s) = \begin{cases} \arg\max_a \mathbb{E}_k[Q_k(s,a)] & \text{if } Var_k[AC_k(s,a)] > Var_{threshold} \\ \pi_{backup}(s) & otherwise \end{cases} \tag{10}$$

4.2 Delta to ID uncertainty estimation

One problem with uncertainty estimation within reinforcement learning is the fact, that there often are alternative decisions to take in an MDP. These can be called alternative possible actions. When an agent is in a state with alternative possible actions, the ensemble may already deviate in its prediction, although trained sufficiently in this state (ID - in-distribution). This means, alternative possible actions will sort of overlay the uncertainty form a lack of training (OOD). Given a overlay of such kind will make it hard to decide, if real uncertainty or only alternative possible actions or both are in place. Therefore, we propose to "compare" the given (potentially OOD) situation to a nearest ID situation (IDD - ID delta), to identify "normal" deviation versus uncertainty deviation. To get a comparison, we propose to subtract ID ACV from the given AVC and use the result as a cleaned version of the ACV for uncertainty indication.

$$Var_{delta}[AC(s,a)] == Var_{OOD}[AC(s,a)] - Var_{ID}[AC(s,a)]. \tag{11}$$

There can be different approaches to get the nearest ID from a given OOD scenario. To simplify, we stick to an OOD scenario with one dedicated OOD Object. We propose two variations as a straight forward approach. At first, the OOD object in the given OOD scenario (observation) will be exchanged with a corresponding ID object - like a known obstacle. At second, the given OOD object will just be removed from the scenario without replacement - or with zero replacement, so to say.

$$Obs_{ID_{obstacle}}(pos_{OODobstacle}) == ID_{obstacle}, \tag{12}$$

$$Obs_{ID_{zero}}(pos_{OODobstacle}) == ID_{zero}. \tag{13}$$

A high delta of the ACV will indicate high uncertainty and a low delta low uncertainty in both cases, respectively. To decide on an uncertain situation in a given state, we propose to use a threshold to mask out insignificant variance-delta to ID. This threshold can be used in future work to switch to a backup policy.

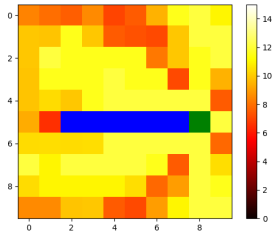
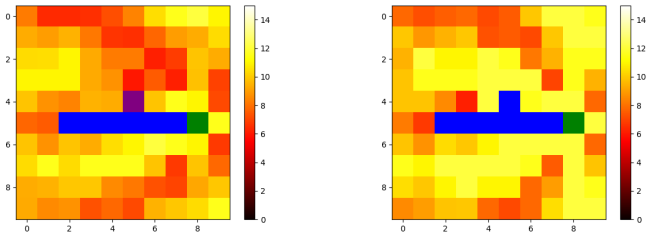


Figure 1: ID Hazard Line as baseline ID scenario.



(a) OOD obstacle at the top (b) ID Hazard at the top

Figure 2: ID Hazard Line with obstacle in the middle

5 Experimental results

5.1 Setup and Training

We trained complete agents in parallel with a random placed set of 10 hazards singly placed in a gridwold of 10x10 positions and training runs of 1 million steps each agent. For testing we set up different scenarios with trained before hazards as ID and added a single dedicated never seen before obstacle as ODD. For the paper we focused exemplary an an ID scenario with a line of known hazards in the middle of the grid and the goal at the end of the line.

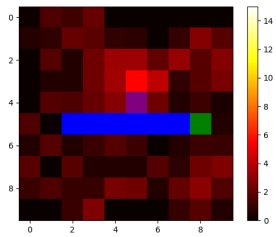
For the visualization of the uncertainty estimation we calculated heatmaps over the grid showing each resulting uncertainty estimation for each position of the agent in the grid given the overall scenario.

5.2 Uncertainty heatmaps

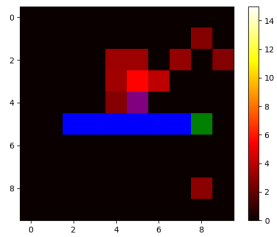
For the depicted results the uncertainty calculation based on the action count variance of the ensemble members is used. Figure 1 is the base scenario with the known ID hazard line in blue and the goal in green. As we use variance in the action count, a higher brighter colour means more concentration on fewer actions (and therefore more certainty) and darker colour means a less concentration in the actions or more equally distributed action (and therefore higher uncertainty. As one can see in the base scenario - due to possible alternative action predictions there are some "uncertainties" along the diagonals to the goal, as these coordinates have equal probabilities vertically and horizontally to approach the goal. This applies e.g. also for the point on the left of the hazard line, as the probabilities for up and down are equally distributed.

Figure 2a shows the predictions with one unknown obstacle inserted in the middle direct on top of the line shown in purple. One can see that in tendency many of the area and especially the area around the unknown obstacle is increased in uncertainty. Nevertheless, the uncertainty indication is superposed by the already given "uncertainty" of the possible alternative predictions from the baseline ID scenario.

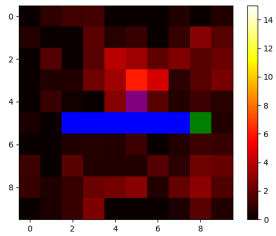
In contrast, figure 2b shows the predictions with a known obstacle inserted in the middle direct on top of the line shown in blue, instead of the OOD obstacle. Now, the uncertainty indication is much more near to the baseline ID scenario.



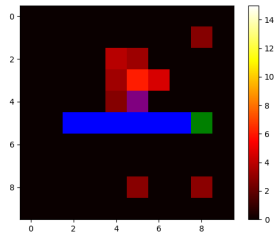
(a) Delta to ID baseline



(b) Delta to ID with threshold



(c) Delta to ID known obstacle



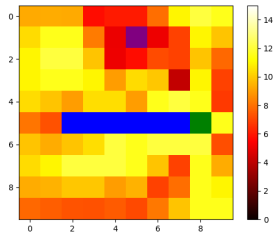
(d) Delta to ID known obstacle with threshold

Figure 3: Delta OOD to ID with obstacle in the middle

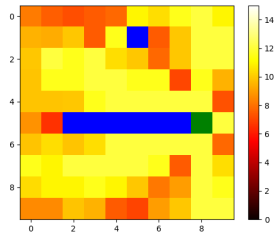
Our idea presented in 4.2 proposes to subtract the baseline variance from the OOD variance and therefore try to eliminate the baseline variance resulting from the possible alternative predictions.

In Figure 3 the results are depicted for the delta to the ID baseline without and with threshold (3a and 3b) and the delta to ID with the known obstacle (3c and 3d). It seems feasible to indicate the given OOD hotspot when masking out a dedicated threshold, like given in 3b and 3d, although the indication is not totally sharp.

In the given scenario, the OOD hotspot is lying direct in an area of low uncertainty (the yellow area on top of the blue line). One can think of a special situation, where the ID subtraction is convenient to isolate the hotspot. Therefore, we also ran setups, where the hotspot is lying in an area of already given "uncertainty" from possible alternative predictions - like in upper middle section. See Figure 4.



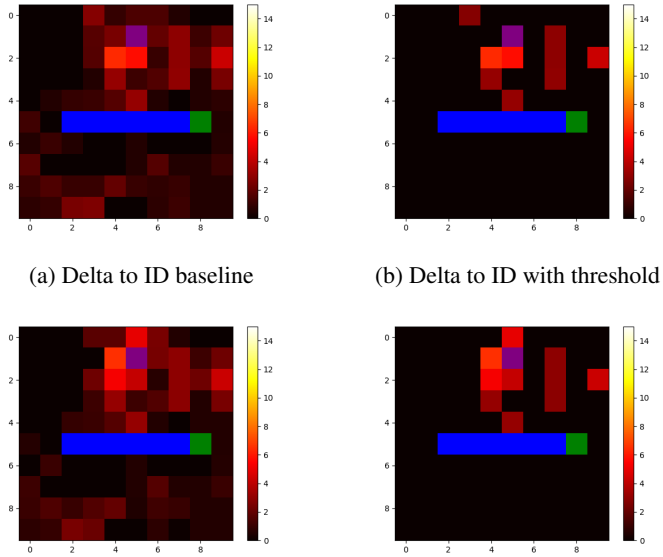
(a) OOD obstacle at the top



(b) ID Hazard at the top

Figure 4: ID Hazard Line with obstacle at the top

And Figure 5 shows the resulting indication for the delta OOD to ID baseline in 5a and 5b - and the delta to ID with known hazard in 5c and 5d. Again, the given threshold seems to isolate the hotspot from the OOD obstacle quite well. But here, the delta to ID baseline fails, as the top and left fields from the obstacle are not indicated. The already higher "uncertainty" form the possible alternative predictions in that area subtracts out the indication. But actually these fields should be indicated - because they lie in the diagonal path to the goal and would therefore be conflicting due to



(a) Delta to ID baseline

(b) Delta to ID with threshold

(c) Delta to ID known obstacle

(d) Delta to ID known obstacle with threshold

Figure 5: a) ID Hazard Line with obstacle at the top

the unknown obstacle. In contrast, the delta to ID with known obstacle replacement performs better and indicates the respective fields.

6 Conclusion

Within Reinforcement Learning for a decision-making agent aleatoric and epistemic uncertainty can be distinguished. Aleatroic uncertainty refers to the inherent randomness of an outcome whereas epistemic uncertainty arises due to lack of knowledge. In this paper we investigated ensemble based epistemic uncertainty estimation on gridworld scenarios with discrete action spaces and overlapping alternative predictions.

We exhibited the problem obstacle with discrete actions spaces and variance calculation and proposed to use action count variance (ACV) instead of action variance. We depicted the conflict of uncertainty estimation due to a lack of training with overlapping alternative solutions. As solution, we proposed to calculate the delta to ID (IDD) for the action count variance and utilized a threshold to indicate uncertainty due to a lack of training, only.

With experiments utilizing representative test scenarios on grid world we could show that action count variance with delta to ID is able to indicate uncertain states based on a threshold calculation with high probability. Therefore, a decision for a backup policy based on that indication seams to be a feasible solution.

In future work we want to investigate feasible methods to determine a sufficient near ID scenario for a given OOD scenario and extend the approach to more general and realistic environments. We also want to focus on strategies to deal with an uncertainty indication and compare the resulting performance to other given approaches.

References

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models. *arXiv:1805.12114 [cs, stat]*, November 2018. arXiv: 1805.12114.

William R. Clements, Bastien Van Delft, Benoît-Marie Robaglia, Reda Bahi Slaoui, and Sébastien Toth. Estimating Risk and Uncertainty in Deep Reinforcement Learning. *arXiv:1905.09638 [cs, stat]*, September 2020. arXiv: 1905.09638.

Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Tom Haider, Felippe Schmoeller Roza, Dirk Eilers, and Karsten Roscher. Domain shifts in reinforcement learning: Identifying disturbances in environments. page 7, 2021.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv:1610.02136 [cs]*, October 2018. arXiv: 1610.02136.

Carl-Johan Hoel, Krister Wolff, and Leo Laine. Tactical Decision-Making in Autonomous Driving by Reinforcement Learning with Uncertainty Estimation. April 2020.

Carl-Johan Hoel, Krister Wolff, and Leo Laine. Ensemble quantile networks: Uncertainty-aware reinforcement learning with applications in autonomous driving. 2021.

Gregory Kahn, Adam Villaflor, Vitchyr Pong, Pieter Abbeel, and Sergey Levine. Uncertainty-Aware Reinforcement Learning for Collision Avoidance. February 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. *arXiv:1807.03888 [cs, stat]*, October 2018. arXiv: 1807.03888.

Björn Lütjens, Michael Everett, and Jonathan P How. Safe reinforcement learning with model uncertainty estimates. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8662–8668. IEEE, 2019.

Sina Mohseni, Mandar Pitale, Jbs Yadawa, and Zhangyang Wang. Self-Supervised Learning for Generalizable Out-of-Distribution Detection. In *AAAI*, 2020.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized Prior Functions for Deep Reinforcement Learning. 2018.

M. Pimentel, D. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Process.*, 2014.

Janis Postels, Hermann Blum, Yannick Strümpler, Cesar Cadena, Roland Siegwart, Luc Van Gool, and Federico Tombari. The Hidden Uncertainty in a Neural Networks Activations. December 2020.

Adrian Schwaiger, Poulami Sinhamahapatra, Jens Gansloser, and Karsten Roscher. Is Uncertainty Quantification in Deep Learning Sufficient for Out-of-Distribution Detection? In *Proc. AISafety@IJCAI2020*, volume 2640 of *CEUR Workshop Proceedings*, page 8, 2020.

Andreas Sedlmeier, Thomas Gabor, Thomy Phan, Lenz Belzner, and Claudia Linnhoff-Popien. Uncertainty-Based Out-of-Distribution Classification in Deep Reinforcement Learning. December 2019.

Felipe Leno Da Silva, Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. Uncertainty-Aware Action Advising for Deep Reinforcement Learning Agents. April 2020.

Pengyue Wang, Yan Li, Shashi Shekhar, and William F. Northrop. Uncertainty Estimation with Distributional Reinforcement Learning for Applications in Intelligent Transportation Systems: A Case Study. October 2019.

Qing Yu and Kiyoharu Aizawa. Unsupervised Out-of-Distribution Detection by Maximum Classifier Discrepancy. *arXiv:1908.04951 [cs]*, August 2019. arXiv: 1908.04951.