Domain Shifts in Reinforcement Learning: Identifying Disturbances in Environments

Tom Haider[1], Felippe Schmoeller Roza[1], Dirk Eilers[1], Karsten Roscher[1], and Stephan Günnemann[2]

Fraunhofer IKS

Fraunhofer Institute for Cognitive Systems IKS

End-to-End Deep Reinforcement Learning (RL) systems return an action no matter what situation they are confronted with, even for situations that differ entirely from those an agent has been trained for. In this work, we propose to formalize the changes in the environment in terms of the Markov Decision Process (MDP), resulting in a more formal framework when dealing with such problems.

Preliminaries

In RL, we consider an agent that sequentially interacts with an environment modeled as a Markov Decision Process (MDP):

1. An MDP is a tuple $\boldsymbol{M} \coloneqq (S, A, R, P, \mu_0)$
2. S is the set of states
3. A is the set of actions
4. $R: S \times A \times S \rightarrow \mathbb{R}$ is the reward function
5. $P: S \times A \times S \rightarrow [0,1]$ is the transition probability function
6. $\mu_0: S \rightarrow [0,1]$ is the starting state distribution

At each timestep the agent observes the current state $s_t \in S$, takes an action $a_t \in A$, transitions to the next state s_{t+1} drawn from the distribution $P(s_t, a_t)$, and receives a reward $R(s_t, a_t, s_{t+1})$.

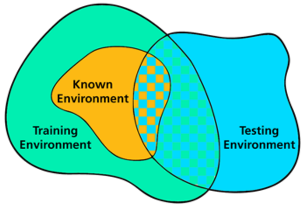
Domain Shift and Related Concepts

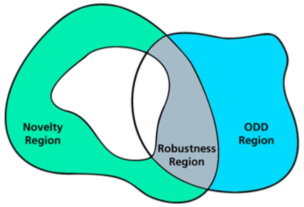
In ML, and RL more specifically, different concepts related to the problem of dealing with changes from the training domain to the testing domain exist.

- Transfer Learning
- Domain Adaptation, and Domain Randomization
- Novelty Detection and Intrinsic Motivation
- Distributional Shift and OOD
- Operational Design Domain (ODD)
- Robust RL

How to map these concepts into RL problems?

- **Training Environment:** all the situations the agent can encounter during training
- **Known Environment:** already explored states
- **Testing Environment:** region the agent can reach after deployment.





- **Novelty Region:** can be explored during training
- **Robustness Region:** novel states the agent has to handle
- **OOD region:** samples from a different distribution

Using these concepts in RL problems is not straightforward and can lead to multiple interpretations! **We rather propose to tackle this problem by decomposing the disturbances in the environment into the elements of the MDP.**

Example scenario

- AGV navigating in a warehouse.
- Goal: reach the destination
- Avoid collisions
- Hazards absent during training

The table in the right side shows how this problem could be decomposed in terms of the MDP.

Decomposition of potential hazards into components of the MDP.

Potential hazards	S	A	P	R	μ_0
Workers interacting with the AGV	√		√		
Other robots interacting with the AGV	√		√		
Changed warehouse layout	√				
Multiple goals				√	
Unusual starting position					√
Malfunctions of the AGV		(√)	√		
Noisy sensors	√				

Each of these hazards can be traced back to only a single or at most two components of the MDP. This mapping is straightforward and directly helps to isolate safety-relevant issues. Once isolated, these issues can be detected and handled more easily.

Conclusion

Decomposing changes in the RL environment into the aspects that build up the MDP allows us to isolate different concerns and treat each of them separately. It is important to notice that the goal is to formally describe these disturbances to help when designing safer systems, while identifying and reacting to such changes is a separate problem. Future work has to detail further how the MDP decomposition and structuring of the problem can be utilized for adding robustness and ensuring safety under environmental disturbances

[1] Fraunhofer IKS
[2] Technical University of Munich