Abstract

Intelligent robots are increasingly being considered for use alongside humans and in critical applications such as healthcare and manufacturing. Reinforcement Learning (RL) can enable agents or robots to learn tasks unsupervised. If RL is to be deployed for real-world robotic applications, the decisions or actions of the RL agent need to be trustworthy. The robustness and explainability of RL need to be improved for RL to be trustworthy in real-world deployments.

The robustness and explainability of RL can be improved by combining RL with a field called Causal Inference. Causal RL is a combination of Causal Inference and RL. Environments are represented as Structural Causal Models (SCM) in Causal RL. The RL agent learns about the underlying SCM through three different interactions: association (seeing), interventions (doing), and counterfactuals (imagining). All three interactions are needed to learn the complete underlying SCM. The causal information that is obtained through interacting with the environment is then stored as a causal representation. The causal representation captures all the invariant causal mechanisms across domains and tasks. The causal representation can be transferred to different tasks with similar causal structures and improves the robustness of RL.

However, currently, there is limited research in Causal RL as it is a recent field. Existing approaches have limitations such as assuming the SCM is known or partially known, the Causal RL solutions are usually not complete, and Causal RL has not been applied to complex tasks like robotics. This thesis proposes the first complete Causal RL solution that is applied to complex robotic tasks. The proposed solution is called Causal Counterfactual (CausalCF). CausalCF learns about the underlying SCM from scratch using the three different Causal RL interactions. The causal knowledge is stored as a causal representation, and the representation is scalable and transferable to different tasks. CausalCF is implemented and evaluated in a realistic robotic simulation environment called CausalWorld. CausalWorld provides a range of complex robotic control and object manipulation tasks.

CausalCF combines ideas from Causal Curiosity and CoPhy. Causal Curiosity provides an approach for using interventions and a causal representation to train an RL agent. CoPhy is a Deep Learning solution that can perform counterfactuals. CoPhy is adapted for use in RL. Each component of the CausalCF design was evaluated in CausalWorld. All the components of the CausalCF design improved the training performance and the robustness of the RL agent. The causal representation learnt from one task was directly transferred to train an RL agent in another task. The causal representations captured the causal mechanisms that remain invariant across tasks and improved the RL agent's training performance and robustness. CausalCF can be directly applied to the different tasks available in CausalWorld.