Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [Explainable Reinforcement Learning Through Casual Lens]
Date: [01-11-2024]

Make sure your critique Address the following points:
1. The problem the paper is trying to address
2. Key contributions of the paper
3. Proposed algorithm/framework
4. How the proposed algorithm addressed the described problem
Note: Be concise with your explanations. Unnecessary verbosity will be penalized. Please don't exceed 2 pages.

---

1 The problem the paper is trying to address

The paper addresses the problem of **explaining the behavior of model-free reinforcement learning (RL) agents** using **causal models**. Specifically, it aims to construct **causal explanations** to answer "why" and "why not" questions about the actions taken by RL agents, enhancing interpretability and trust in AI systems through a structured **causal framework** that includes:

1. **Structural Causal Models (SCMs)** to represent causal relationships between actions and outcomes.

2. **Action Influence Models** for defining and quantifying the causal impact of agent actions.

This approach enables generation of explanations in terms of counterfactuals, which improves users' understanding and prediction of agent behavior in complex environments.

2 Key contributions of the paper

- Introduction and formalization of an **Action Influence Model** for model-free reinforcement learning (RL) agents, utilizing **Structural Causal Models (SCMs)** to capture causal relationships between actions and outcomes.
- Development of a method to generate **contrastive explanations** (i.e., explanations for "why" and "why not" questions) through **counterfactual analysis** of the SCM, enabling causal-based explanations for actions taken.
- **Computational evaluation** of the proposed model across 6 RL domains to measure its performance in task prediction and explanation accuracy.
- Conducting a **human study** with 120 participants to assess the impact of causal explanations on task prediction accuracy, explanation satisfaction, and trust in the agent's behavior.

3 Proposed algorithm/framework

- **Define Causal Relationships:** Construct an **Action Influence Model** by specifying causal relationships between actions and state variables using **Structural Causal Models (SCMs)**.
- **Learn Structural Equations:** During RL training, learn the structural equations for each causal relationship, using data from the environment to approximate the effects of each action.
- **Generate Explanations:**
  - Identify the **causal chain** leading from the action to the outcome.
  - Generate **counterfactual explanations** by comparing the causal chains of the observed action with those of alternative actions.
- **Evaluate and Validate:** Use the trained model to predict task actions in test environments and validate explanation quality through human studies.

4 How the proposed algorithm addressed the problem

- **Modeling Causal Influence:** By constructing an **Action Influence Model** based on **Structural Causal Models (SCMs)**, it captures causal relationships between actions and outcomes in reinforcement learning (RL) agents, making it possible to explain why specific actions lead to observed results.
- **Generating Counterfactual Explanations:** For "why" and "why not" questions, the framework generates explanations by contrasting the causal effect of the chosen action with that of alternative actions, offering an intuitive understanding of the agent's behavior through **counterfactual reasoning**.
- **Enhancing Transparency and Predictability:** The algorithm allows users to predict the agent's future actions and assess its decisions in a transparent way, fostering trust by providing detailed, causal-based answers to user queries about agent behavior.
- **Improving Human Interpretability:** Through computational evaluation and human studies, the algorithm demonstrates improved user comprehension, satisfaction, and accuracy in predicting agent actions, addressing the interpretability gap in model-free RL by aligning explanations with human reasoning.