**Explainable Reinforcement Learning for Broad-XAI: A Conceptual Framework and Survey**

[**https://arxiv.org/pdf/2108.09003.pdf**](https://arxiv.org/pdf/2108.09003.pdf)

1. **INTRODUCTION**

|  |
| --- |
| 이 논문은 다음과 같이 구성되어 있다.   * 1. XRL (eXplainable Reinforcement Learning)의 **배경 및 XRL이 다루는 영역** * 2. XRL의 **conceptual framework** * 3. 프레임워크의 **initial stage**에 대한 현재의 접근 방식 * 4. 프레임워크의 이후 단계에 대한 **연구 기회** * 5. 프레임워크가 **broad-XAI의 개발에 어떻게 융합**되어 이용할 수 있는가? |

1. **Levels of Explanation for XAI**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 왼쪽 그림은 **XAI의 Explanation Level**을 나타낸 것이다.  여기서 각 level은 다음을 나타낸다.  (Dazeley et al.’s meta-survey identifies diverse subfields of XAI research [58])   |  |  | | --- | --- | | Zero-order Explanations | Agent의 **즉각적인 행동**   * 의도를 반영하지 않음 | | First-order Explanations | **현재 목표**에 기반한, agent의 행동에 대한 설명   * Agent의 현재 disposition에 기반 | | Second-order Explanations | **사회적인 상황 또는 인식**에 기반한, agent의 행동에 대한 설명   * 같은 환경에 있는 다른 actor를 고려한 설명 | | N-th … | Agent가 **cultural expectation**에 의해 **어떻게 행동을 바꾸는지**에 대한 설명 | | Meta explanations | 설명을 생성할 때의 프로세스에 대한 설명 | |

1. **Extensions of Basic Reinforcement Learning (RL) approaches**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[1] Growing state space**  **State space가 커져서 Agent의 search 영역의 복잡도가 커지는** 문제이다.   * 이때, Neural Network와 같은 function approximation 기술이 많이 쓰인다. * 이것은 Deep RL (DRL)을 포함한다.   **[2]** **Multi-goal RL**  Reinforcement Learning의 **목표(goal)가 여러 개일 때**는, agent는 이 목표를 성취해야 할 뿐만 아니라, **적절한 sub-goal**을 설정할 수 있어야 한다.   * 이때, 최종적인 goal은 agent의 궁극적인 목표를 나타낸다. * **Multi-objective RL (MORL)**은 primary objective와 균형이 맞춰져야 하는 **conflicting objectives**가 존재할 수 있을 것으로 간주한다.   **[3] Difference between RL and supervised learning**   |  |  |  | | --- | --- | --- | |  | Supervised Learning | Reinforcement Learning | | learning | **각각의 입력을 출력으로 individual**하게 mapping시킨다.   * XAI에서 **input component** 또는 결과에 해당하는 분류를 생성하기 위한 **process stage를 identify**하면 된다. * 각 instance는 **stand-alone instance**이며, local explanation은 이 사실에 기반한다. | **각 instance 간에 implicit한 관계**가 존재한다.   * 이것은 next state가 **previous state에서 실행된 행동**에 의해서 visit되기 때문이다. * 이것은 state, action, 그리고 다음 state 간의 **temporal한 의존 관계**를 생성한다. | | mapping | **Right answer**를 얻기 위한 mapping을 한다. 즉, classification 또는 regression 값을 제공하기 위해 **학습된 mapping을 사용**한다. | **Reward signal을 최대화**하는 것에 중점을 둔다.   * Reward는 agent의 목표를 나타낸다. |   Reinforcement learning과 supervised learning의 이러한 근본적인 차이점은 설명 방법을 **단순한 해석 방법과는 조금 다르게** 생각해야 하게 한다.   * RL은 잠재적으로 **broad-XAI 시스템을 개발**할 수 있게 한다. * 이 프레임워크의 context에 기반하여, 이 논문은 **설명을 제공하기 위한 현재의 시도** 및 broad-XAI의 개발에 도움이 되는 **잠재적인 접근 방법**에 대해 연구한다. |

1. **Conceptual Framework for Explainable Reinforcement Learning (XRL)**

|  |
| --- |
| 왼쪽 그림은 **Causal Explanation Network (CEN) 모델**의 reproduction을 나타낸 것이다.  (Bohm and Pfister [28])  각 node는 **사람의 행동을 설명할 때 사용되는 component**이고, node 간의 arc는 사람이 설명을 제공할 때의 이들 개념에 대한 causal link를 나타낸다.   * 반드시 outcome에 대한 full and direct한 순서를 반영할 필요는 없지만, **사람들이 전형적으로 사용하는 causal explanation**을 나타낸다.     왼쪽 그림은 Explainable Reinforcement Learning의 **conception framework**을 나타낸다.   * 위 그림에서 **같은 causal pathway를 설명에 사용할 수 있도록** 한 것이다. * 각 node는 **agent가 행동을 결정할 때 사용되는 process**를 나타낸다. * 각 edge (joining node)는 **agent의 행동을 일반화**할 때 사용될 수 있는 **causal relationship**을 나타낸다.   Reinforcement Learning에서는 **agent가 특정한 goal 또는 objective를 추구**할 때 action이 발생하며, 이것은 single goal 또는 goal의 hierarchy이다.   * 각각은 agent의 설명을 생성하기 위해 **cycle**될 수 있다. * Goal은 single objective 또는 서로 균형을 맞춰야 하는 multiple objective이다.   + 이때, agent는 **이들 goal/objective 간에 switch**를 할 수 있는데, 그 이유는 우선 순위의 internal change 또는 더 큰 목표를 해결하는 진행 상황 등이다. |
| (next page) |

1. **Conceptual Framework for Explainable Reinforcement Learning (XRL) (Cont.)**

|  |
| --- |
| (previous page) |
| 위 그림은 **Reinforcement Learning에 대한 전통적인 접근 방식**을 사용했을 때 설명 가능한 행동들의 종류를 나타낸 것이다.   * **전통적인 RL에서의 causal explanation**을 나타난다. * **이 모델은 agent에 의해 인식되는 것 + agent의 행동만을 포함한다.** * 이 프레임워크는 preception과 action 간의 causal link 중 **전체 모델에는 없는 것을 일부 포함**한다.   + 이 링크들은 **더 깊은 causal path를 나타내는 behavioural component들을 대체**하며, Simplified-CXF를 이용하여 모델링/설명되지 않는다고 가정한다.   이 모델에서는 agent는 **하나의 preset goal**을 가지고 있으며, **objective는 이 goal에 대한 reward를 최대화**하는 것이라고 가정한다.   * 이러한 시뮬레이션에서, goal은 사용자에게 알려져 있거나, 또는 긴 시간 동안의 행동을 통해서 관측될 수 있다.   Goal/objective, disposition 또는 expectation과 event를 제거했을 때, agent는 이 모델의 causal path를 이용할 수 없다. 따라서, 이 프레임워크는 **이러한 행동을 skip할 수 있게 하는, perception to action의 causal path**를 반드시 포함해야 한다.   * 이 causal path는 **전체 프레임워크의 part가 아니다.**   Agent의 perception에 대한 설명은 다음 중 1가지 이상을 자세히 설명하는 것을 목표로 한다.   |  |  | | --- | --- | | Perception | Agent가 **현재 환경을 어떻게 인식**하였는가? | | Introspective | 인식한 state가 **선택된 action에 어떻게 기여**했는가? | | Contrastive | 인식한 state가 왜 **다른 action**을 선택하게 하지 않았는가? | | Counterfactual | Agent가 다른 action을 선택하기 위해서, **perception의 change** 중 무엇이 필요한가? | | Influenced | 인식된 state가 **최종 출력**에 어떻게 영향을 미쳤는가? | |

1. **Types of interpretation that can be generated from an interpretable ML model**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 왼쪽 그림은 **interpretable한 기계학습 모델**에 의해 생성될 수 있는 interpretation의 종류를 나타낸다.   |  |  | | --- | --- | | Feature summary | **Statistics 또는 visualizations**를 사용하여, **출력에 도달**했을 때 중요도가 가장 높은 feature들과 그것들 사이의 관계를 보여준다. | | Representation of the internal model’s operation  (Model Internals) | **Internal model의 연산**을 표현한다.   * fired rule 또는 neuron, 또는 evaluation process를 통과하는 pathway가 있다. | | Data point | **유사하거나 관련된 data point**를 identify한다.   * 같은 class의 이미지 등 | | Intrainsically interpretable model | **Secondary intrinsically interpretable model의 구성**을 이용한다.   * 앞서 언급된 interpretation method 중 한 가지를 이용한다. | |

1. **Explanations of actions**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Agent의 행동에 대한 설명은 다음을 detail하게 설명하는 것을 목표로 한다.   |  |  | | --- | --- | | Introspective | **Action이 왜 선택**되었는가? | | Contrastive | **다른 action**이 왜 선택되지 않았는가? | | Influenced | 실행된 action이 **최종 출력에 어떻게 영향**을 미쳤는가? | | Counterfactual | **이전의 어떤 행동**이 특정한 alternative action이 선택되도록 하였는가? | |

1. **Conversational model for explanation**

|  |
| --- |
| 위 그림은 **explainee가 만족할 때까지 agent가 3개의 단계를 반복**할 때의 conversational model이다.  Agent는 **explanation의 가장 높은 단계**에서 시작하며, 하위 단계로 내려간다. |