Awesome Explainable Reinforcement Learning

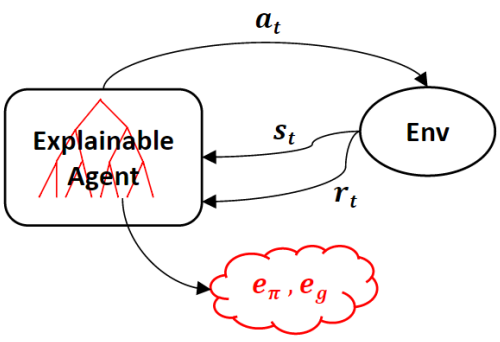
A list of awesome paper and code resources on explainable reinforcement learning (XRL). Inspired by Awesome-Visual-Transformer, awesome-transfer-learning, awesome-self-supervised-learning, awesome-graph-self-supervised-learning and awesome-deep-vision.

Table of Contents

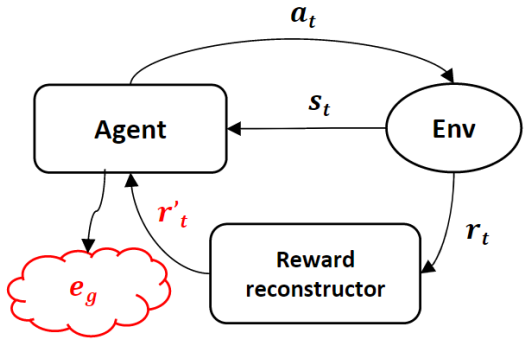
- Overview
- Surveys
- Explainability in RL
  - Model Explaining
  - Reward Explaining
  - State Explaining
  - Task Explaining
- Human knowledge for RL paradigm

Overview

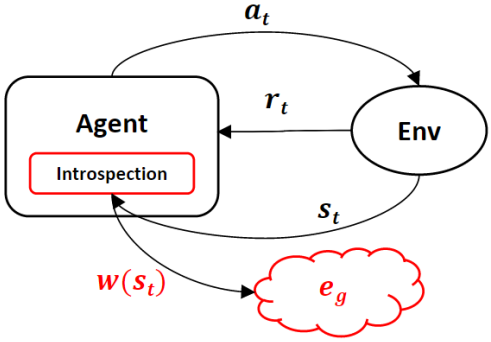
We review current explainable reinforcement learning framework and explainability of human knowledge-based reinforcement learning framework. We creatively propose a new taxonomy for existing explainable reinforcement learning based on reinforcement learning paradigm. Specifically, we divide existing explainable reinforcement learning methods into four categories: model-explaining, reward-explaining, state-explaining, task-explaining as shown below:



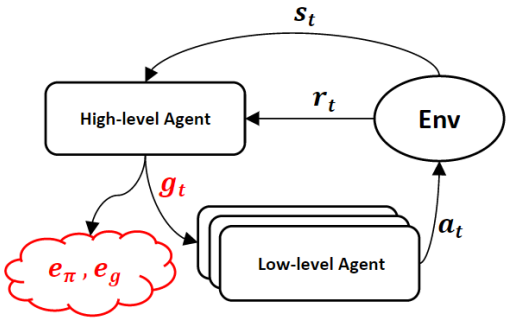
(a) Model Explaining



(b) Reward Explaining



(c) State Explaining



(d) Task Explaining

In the Figure above, e_π and e_g denotes for two aspects of explanations: inner logic inference of agent and goal influence in action-taking. a_t, r_t, s_t refer to the action, reward and states at time t; The red parts in these diagrams are the explainable parts.

- Model Explaining: trains the agent to be explainable by having understandable logic operation in its inner structure
- Reward Explaining: reconstructs reward function towards an explainable one r_t' and makes it possible to see how the goal influences the agent
- State Explaining: adds a submodule for introspection to quantify the influences of different state features towards the decision-making as $w(s_t)$.
- Task Explaining: gets an architectural level explainability in complex environments by multilevel agents, the high-level agent schedule low-level agents by the subgoal signal g_t which could be utilized for explanations.

To know more about existing XRL framework and our taxonomy, the existing XRL papers within different typs are listed below and are categorize into our taxonomy. For each paper, we provide the soure file and the project code if the paper has source code and the code is pubilic.

Surveys

- Explainable Reinforcement Learning: A Survey
  - E. Puiutta and E. Veith. *CD-MAKE 2020*. [paper]
- A Survey on Interpretable Reinforcement Learning
  - C. Glanois, P. Weng, M. Zimmer, D. Li, T. Yang, J. Hao and W. Liu. *Arxiv 2021*. [paper]
- Explainable Reinforcement Learning for Broad-XAI: A Conceptual Framework and Survey
  - R. Dazeley, P. Vamplew and F.Cruz. *Arxiv 2021*. [paper]
- Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends
  - Lindsay Wells and Tomasz Bednarz. *FRAI 2021*. [paper]
- Explainability in deep reinforcement learning
  - A. Heuillet, F. Couthouis and N. Díaz-Rodríguez. *KBS 2021*. [paper]
- Explainable Deep Reinforcement Learning: State of the Art and Challenges
  - G. Vouros. *ACM 2022*. [paper]

Explainability in RL

Model Explaining

- Explainable Reinforcement Learning: A Survey
  - E. Puiutta and E. Veith. *CD-MAKE 2020*. [paper]
- A Survey on Interpretable Reinforcement Learning
  - C. Glanois, P. Weng, M. Zimmer, D. Li, T. Yang, J. Hao and W. Liu. *Arxiv 2021*. [paper]
- Explainable Reinforcement Learning for Broad-XAI: A Conceptual Framework and Survey
  - R. Dazeley, P. Vamplew and F.Cruz. *Arxiv 2021*. [paper]
- Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends
  - Lindsay Wells and Tomasz Bednarz. *FRAI 2021*. [paper]

- Explainability in deep reinforcement learning
    - A. Heuillet, F. Couthouis and N. Díaz-Rodríguez. *KBS 2021*. [paper]
- Explainable Deep Reinforcement Learning: State of the Art and Challenges
    - G. Vouros. *ACM 2022*. [paper]

Reward Explaining

- Explainable Reinforcement Learning: A Survey
    - E. Puiutta and E. Veith. *CD-MAKE 2020*. [paper]
- A Survey on Interpretable Reinforcement Learning
    - C. Glanois, P. Weng, M. Zimmer, D. Li, T. Yang, J. Hao and W. Liu. *Arxiv 2021*. [paper]
- Explainable Reinforcement Learning for Broad-XAI: A Conceptual Framework and Survey
    - R. Dazeley, P. Vamplew and F.Cruz. *Arxiv 2021*. [paper]
- Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends
    - Lindsay Wells and Tomasz Bednarz. *FRAI 2021*. [paper]
- Explainability in deep reinforcement learning
    - A. Heuillet, F. Couthouis and N. Díaz-Rodríguez. *KBS 2021*. [paper]
- Explainable Deep Reinforcement Learning: State of the Art and Challenges
    - G. Vouros. *ACM 2022*. [paper]

State Explaining

Task Explaining

Human knowledge for RL paradigm