Proposal: Knowledge-Graphs for Feature Importance Explainability in Deep Reinforcement Learning Models

Roel Leenders
University of Twente, Enschede, The Netherlands
r.leenders@student.utwente.nl

1 Introduction

In recent years the field of Artificial Intelligence (AI) has seen an increasing need for explainability. As models became more complex and computationally intensive, a performance-transparency trade-off was introduced (Puiutta and Veith, 2020). High-performing models with complex inner workings come at the cost of transparency; it becomes less clear how they achieve their decisions and predictions.

The necessity for explainable AI (XAI) increases particularly for machine learning methods like Reinforcement Learning (RL) where an agent learns autonomously with little to no human intervention. Additionally, since people are by and large concerned about the risks of automated decision-making (Araujo et al., 2020), explainability is also important to improve the public opinion and trust of AI.

2 Problem Statement

Fortunately, XAI has seen a rapid growth in active research (Tiddi and Schlobach, 2022a) with a plethora of contributions proposing a variety of different techniques (Xu et al., 2015; Ribeiro et al., 2018; Byrne, 2019). Although still understudied, RL has seen emerging XAI trends (Wells and Bednarz, 2021) which include visualization, query-based explanations and policy summarization. An XAI method that makes models inherently more transparent is the use of a knowledge-driven (symbolic) methods (Tiddi and Schlobach, 2022b). Oltramari et al. (2020) propose a method in which knowledge-driven methods (e.g. knowledge-graphs) can be used in hybrid fashion with deep neural networks. The authors show that this neuro-symbolic approach is able to maintain interpretability while achieving comparable performance. However, since knowledge-driven methods for XAI are more common within the supervised learning literature (Tiddi and Schlobach, 2022b), knowledge-driven XAI used for RL is still very much understudied. Therefore, the proposed research focuses on the use of domain specific knowledge graphs to both train and explain the behavior of a RL agent. This results in the following research question and sub-questions:

- **RQ. 1**: To what extent, if any, can a domain specific knowledge-graph improve the feature importance explainability of a deep RL model?
- **SQ. 1**: How
- **SQ. 2**: How

3 Proposed Method of Research

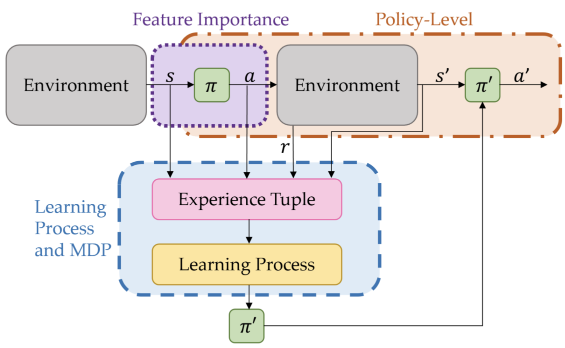

Figure 1: RL taxonomy proposed by Milani et al. (2022) which distinguishes between *feature importance*, *learning process and MDP*, and *policy-level* XAI methods.

To narrow the scope of the proposed research, the focus will mainly be on the *feature importance* explainability of a RL model (Milani et al., 2022). Milani et al. propose a taxonomy for organizing the XAI literature that focuses on RL. This taxonomy distinguishes between three different aspects of explainable RL (see figure 1). In particular, *feature importance* aims to explain the features that affect an agent's decision-making for a given input state.

Most *feature importance* techniques mentioned in the taxonomy use XAI methods extended from supervised learning literature (Greydanus et al., 2018; Goel et al., 2018; Ehsan et al., 2017). Since there is supporting literature for knowledge-driven XAI in supervised learning, the proposed research will aim to use knowledge-graphs for *feature importance* explainability in deep RL models.

4 Planning

References

Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. 2020. In ai we trust? perceptions about automated decision-making by artificial intelligence. *AI Soc.*, 35(3):611–623.

Ruth M. J. Byrne. 2019. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6276–6282. International Joint Conferences on Artificial Intelligence Organization.

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2017. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. Technical Report arXiv:1702.07826, arXiv. ArXiv:1702.07826 [cs] type: article.

Vik Goel, Jameson Weng, and Pascal Poupart. 2018. Unsupervised Video Object Segmentation for Deep Reinforcement Learning. Technical Report arXiv:1805.07780, arXiv. ArXiv:1805.07780 [cs] type: article.

Sam Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. 2018. Visualizing and Understanding Atari Agents. Technical Report arXiv:1711.00138, arXiv. ArXiv:1711.00138 [cs] type: article.

Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2022. A survey of explainable reinforcement learning.

Alessandro Oltramari, Jonathan Francis, Cory Henson, Kaixin Ma, and Ruwan Wickramarachchi. 2020. Neuro-symbolic architectures for context understanding.

Erika Puiutta and Eric MSP Veith. 2020. Explainable reinforcement learning: A survey.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Ilaria Tiddi and Stefan Schlobach. 2022a. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, 302:103627.

Ilaria Tiddi and Stefan Schlobach. 2022b. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, 302:103627.

Lindsay Wells and Tomasz Bednarz. 2021. Explainable ai and reinforcement learning—a systematic review of current approaches and trends. *Frontiers in Artificial Intelligence*, 4.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention.