Szkoła Doktorska Technologii Informacyjnych i Biomedycznych
Polskiej Akademii Nauk (TIB PAN)

TEMAT:

Interpretability of Deep Reinforcement Learning Agents
Interpretowalność agentów głębokiego uczenia się ze wzmocnieniem

PROMOTOR:

dr hab. inż. Przemysław Sekuła
psekula@iitis.pl
Instytut Informatyki Teoretycznej i Stosowanej PAN
Gliwice, Bałtycka 5

OPIS:

Deep Reinforcement Learning (Deep RL) agents have demonstrated remarkable capabilities, achieving superhuman performance in complex sequential decision-making tasks such as mastering Atari games. However, despite their success, these agents often function as opaque "black boxes," making their internal decision-making processes difficult to understand. This fundamental lack of transparency presents significant barriers, hindering user trust and limiting the safe deployment of these powerful systems in critical real-world applications.

Addressing this challenge, this research focuses on developing and evaluating novel methods specifically designed to enhance the interpretability of Deep RL agents. The central aim is not merely to observe what actions an agent takes, but to explain why it chooses a particular action in a given situation, thereby uncovering the underlying rationale driving its behavior. Potential avenues for achieving this include techniques such as analyzing network activations via saliency maps, approximating complex neural network policies with simpler, inherently interpretable models, or methods that decompose the agent's reward signals to attribute value to different factors.

To ensure rigor and relevance, the effectiveness of these developed explanation methods will be systematically tested and validated using well-established Deep RL benchmarks, such as the Atari game suite. Success will be measured by the fidelity of the explanations—how accurately they reflect the agent's true reasoning—as well as their practical utility for human understanding and identifying potential flaws or biases in agent behavior.

Ultimately, this work seeks to contribute new techniques and robust frameworks tailored for explaining sequential decision-making in the context of Deep RL. By enhancing interpretability, the research aims to foster greater trust in autonomous systems and facilitate the responsible design, development, and deployment of more understandable and reliable AI agents operating within complex environments.

Nanowsze rozwiązania i techniki Deep RL osiągają w złożonych sekwencyjnych zadaniach podejmowania decyzji, takich jak opanowanie gier Atari wydajność przekraczającą wyniki osiągane przez ludzi. Jednak pomimo swojego sukcesu, rozwiązanie te często działają jak nieprzejrzyste „czarne skrzynki", co utrudnia zrozumienie ich wewnętrznych procesów podejmowania decyzji. Ten fundamentalny brak przejrzystości stanowi poważną barierę, utrudniając zaufanie użytkowników i ograniczając bezpieczne wdrażanie systemów Deep RL w krytycznych aplikacjach w świecie rzeczywistym. Proponowana praca doktorska powinna sprostać temu wyzwaniu, koncentrując się na opracowaniu i ocenie nowych metod, zaprojektowanych w celu zwiększenia interpretowalności Deep RL. Głównym celem jest uzyskanie zdolności do wyjaśnienia, dlaczego Agent wybiera określone działanie w danej sytuacji, odkrywając w ten sposób podstawową logikę kierującą jego zachowaniem. Potencjalne sposoby osiągnięcia tego celu obejmują techniki takie jak analiza aktywacji sieci za pomocą saliency maps, aproksymacja złożonych zasad sieci neuronowych za pomocą prostszych, wrodzonych modeli interpretowalnych lub metody rozkładające sygnały nagrody agenta w celu przypisania wartości różnym czynnikom.

W celu zapewnienia trafności, skuteczności opracowanych metod, planuje się wytestowanie ich użyciu dobrze ugruntowanych testów porównawczych Deep RL, takich jak Atari Games. Sukces będzie mierzony wiernością wyjaśnień — jak dokładnie odzwierciedlają one prawdziwe rozumowanie agenta — a także ich praktyczną użytecznością dla ludzkiego zrozumienia i identyfikowania potencjalnych wad lub uprzedzeń w zachowaniu agenta.

Podsumowując – proponowany temat pracy doktorskiej skupia się na zaproponowaniu nowych technik i frameworków dostosowanych do wyjaśniania podejmowania decyzji w kontekście Deep RL. Poprzez zwiększenie interpretowalności badania mają na celu wzbudzenie większego zaufania do systemów autonomicznych i ułatwienie odpowiedzialnego projektowania, rozwoju i wdrażania bardziej zrozumiałych i niezawodnych agentów AI działających w złożonym otoczeniu.

BIBLIOGRAFIA:

- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). **Human-level control through deep reinforcement learning**. *Nature*, 518(7540), 529-533. https://www.nature.com/articles/nature14236
- Greydanus, S., Koul, A., Dodge, J., & Fern, A. (2018). **Visualizing and Understanding Atari Agents**. arXiv:1711.00138. https://arxiv.org/abs/1711.00138
- Atrey, A., Clary, K., & Jensen, D. (2020). **Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning**. arXiv:1912.05743. https://arxiv.org/abs/1912.05743
- Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., & Doshi-Velez, F. (2019). **Explainable Reinforcement Learning via Reward Decomposition**. IJCAI XAI Workshop. https://web.engr.oregonstate.edu/~afern/papers/reward_decomposition__workshop_final.pdf
- Arjona-Medina, J. A., Gillhofer, M., Widrich, M., Unterthiner, T., Brandstetter, J., & Hochreiter, S. (2019). **RUDDER: Return Decomposition for Delayed Rewards**. *NeurIPS 2019*. https://arxiv.org/abs/1806.07857
- Lin, Z., Zhao, L., Yang, D., Qin, T., Yang, G., & Liu, T.-Y. (2019). **Distributional Reward Decomposition for Reinforcement Learning**. *NeurIPS 2019*. https://arxiv.org/abs/1911.02166
- Huber, T., Schiller, D., & André, E. (2019). **Enhancing Explainability of Deep Reinforcement Learning Through Selective Layer-Wise Relevance Propagation**. *KI 2019: Advances in Artificial Intelligence*. https://link.springer.com/chapter/10.1007/978-3-030-30179-8_16
- Guo, S., Zhang, R., Liu, B., Zhu, Y., Ballard, D., Hayhoe, M., & Stone, P. (2021). **Machine versus Human Attention in Deep Reinforcement Learning Tasks**. *NeurIPS 2021*. https://arxiv.org/abs/2010.15942
- Xing, J., Nagata, T., Zou, X., Neftci, E., & Krichmar, J. L. (2022). **Policy Distillation with Selective Input Gradient Regularization for Efficient Interpretability**. arXiv:2205.08685. https://arxiv.org/abs/2205.08685
- Lu, W., Zhao, X., Fryen, T., Lee, J. H., Li, M., Magg, S., & Wermter, S. (2024). **Causal State Distillation for Explainable Reinforcement Learning**. *CLeaR 2024*. https://arxiv.org/abs/2401.00104
- Delfosse, Q., Sztwiertnia, S., Rothermel, M., Stammer, W., & Kersting, K. (2024). **Interpretable Concept Bottlenecks to Align Reinforcement Learning Agents**. arXiv:2401.05821. https://arxiv.org/abs/2401.05821
- Gokhale, G., Karimi Madahi, S. S., Claessens, B., & Develder, C. (2024). **Distill2Explain: Differentiable Decision Trees for Explainable Reinforcement Learning in Energy Application Controllers**. arXiv:2403.11907. https://arxiv.org/abs/2403.11907
- Kohler, H., Delfosse, Q., Akrour, R., Kersting, K., & Preux, P. (2024). **Interpretable and Editable Programmatic Tree Policies for Reinforcement Learning**. arXiv:2405.14956. https://arxiv.org/abs/2405.14956