AI533 Intelligent Agent and Decision Making
Winter 2024

Term Paper

Paper1. Discovering Reinforcement Learning Algorithms [NeurIPS 2020]

Authors: Junhyuk Oh, Matteo Hessel, Wojciech M. Czarnecki, Zhongwen Xu, Hado van Hasselt, Satinder Singh, David Silver

1.1 Key problem addressed

The key problem addressed is a new approach of automating the discovery of RL algorithms. Specifically, it explores whether it is feasible to discover alternatives to fundamental RL concepts such as value functions and temporal-difference learning entirely from scratch, through data, which includes a system that can learn both "what to predict" such as value functions and "how to learn from it" like bootstrapping, by interacting with a variety of environments. The goal is to develop more efficient algorithms or those that are better adapted to specific environments without relying on manually discovered update rules.

1.2 Summarize

The paper is about create meta-learning framework aimed at automating the discovery of RL algorithms, marking a significant stride in questioning and potentially replacing core RL concepts like value functions and temporal-difference learning with data derived alternatives. The authors present the Learned Policy Gradient (LPG), a method that innovatively learns the entirety of RL update rule, which encompasses both the identification of critical predictive elements and the application of these insights to policy refinement, through varied environment interactions. This approach not only demonstrates LPG’s skills in devising novel equivalents to traditional value functions and effectively utilizing these predictions but also showcases its remarkable ability to scale from simple experimental setups to complex real-world scenarios like Atari games. The result shows the efficiency and adaptability of LPG, suggesting that a shift towards data-driven RL algorithm discovery could yield more dynamic and universally applicable solutions, moving beyond the confines of conventional and manually crafted methods.

1.3 Strengths & Weaknesses

The paper presents a novel approach to the discovery of reinforcement learning (RL) algorithms through the Learned Policy Gradient (LPG), marking a significant step towards automating RL algorithm design. This approach not only demonstrates an ability to generalize from simple environments to complex scenarios like Atari games, suggesting broad applicability, but also reveals some areas for improvement. While the paper makes an empirical advance, it falls short on providing deep theoretical insights into the mechanics behind LPG's success, leaving questions about its foundational principles and conditions for optimal performance. Additionally, the formulation of the problem could benefit from a more thorough exploration of potential limitations and scenarios where the approach might not be as effective. The lack of comparative analysis with other meta-learning or algorithm discovery methods in the experiments leaves a gap in fully understanding LPG's relative performance, including whether it offers a faster or more efficient solution compared to existing algorithms. Despite these weaknesses, the introduction of a meta-learning framework for RL algorithm discovery stands out for its novelty, though its significance could be amplified by addressing the mentioned shortcomings, particularly by enriching the theoretical foundation and enhancing the experimental comparisons.

1.4 Experiment & Results

The experiments are designed to test the LPG's ability to discover useful prediction semantics and evaluate its performance in both training environments and on unseen complex tasks like Atari games. The training involves different kinds of toy domains, including tabular grid worlds and delayed chain MDPs, to capture basic RL challenges. The results validate the main claims, showing that LPG can discover useful functions akin to value functions, effectively utilize them through bootstrapping, and generalize to perform competently on complex Atari games, even outperforming humans in some instances. This demonstrates the framework's potential in discovering general-purpose RL algorithms and supports the hypothesis that data-driven discovery of RL algorithms is feasible and promising.

1.5 Key takeaways & Limitations

The framework they suggest not only demonstrates an impressive ability to generalize from simple, toy environments to complex scenarios like Atari games, suggesting the potential for developing versatile, general-purpose RL algorithms but also successfully identifies alternatives to fundamental RL concepts such as value functions and bootstrapping mechanisms. Despite its empirical successes and the groundbreaking steps toward RL algorithm automation, the paper acknowledges limitations, including a lack of deep theoretical underpinning, unexplored edge cases, reliance on the diversity of training environments for generalization, limited comparative analysis with other methods, and the practicality concerns due to the computational resources required. These aspects underscore the balance between LPG's innovative contributions to the field and the avenues left to explore in future research.

Paper2. Multi-Objective Reinforcement Learning for Designing Ethical Environments [IJCAI 2021]

Authors: Manel Rodriguez-Soto, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar

2.1 Key Problem Addressed

The key problem addressed in this paper is ensuring that autonomous agents learn to behave ethically, aligning their actions with moral values, within the context of RL environments. Specifically, the authors tackle the challenge of designing environments that not only encourage agents to pursue their individual objectives but also guarantee that they adhere to ethical principles during the learning process. This involves formalizing and solving the Ethical Embedding Problem within a multi-objective reinforcement learning framework to ensure that an agent's behavior is ethically optimal while pursuing its goals.

2.2 Summarize

The paper tackles the challenge of ensuring that autonomous agents behave ethically by proposing a novel methodology within the Multi-Objective Reinforcement Learning (MORL) framework. It introduces a theoretical foundation for the Ethical Embedding Problem, allowing for the integration of agent's objectives with ethical guidelines, and develops an algorithm for creating learning environments that theoretically guarantee ethical behavior. The approach formalizes ethical behavior within Multi-Objective Markov Decision Processes (MOMDPs), identifies ethical-optimal policies, and utilizes an algorithm to design environments where ethical learning is assured. The paper’s significant contributions include providing theoretical guarantees for ethical behavior in autonomous agents and validating these through the Public Civility Game, showing the algorithm’s ability to encourage ethical actions like proper garbage disposal while fulfilling individual goals. This work advances the field of machine ethics by integrating ethical considerations into the reinforcement learning design process, offering a practical framework for developing morally aligned AI systems.

2.3 Strengths & Weaknesses

The paper suggested for its theoretical innovation. It offers a novel MORL framework that formalizes the guarantee of ethical behavior in autonomous agents, thus addressing a critical gap in AI ethics. Its sound formulation of the Ethical Embedding Problem elegantly integrates ethical considerations with an agent's objectives. In contrast, the development of an algorithm providing theoretical assurances for ethical learning marks a significant empirical advance. Despite these strengths, the paper's limited empirical validation, particularly its reliance on the Public Civility Game example without extensive testing in diverse scenarios, raises questions about its robustness and scalability. Additionally, the lack of comparison with existing methods and a detailed discussion on handling edge cases or adapting to evolving ethical standards indicates potential areas for improvement.

2.4 Experiment & Results

The experiment setup in the paper is centered around the Public Civility Game, a simplified scenario designed to test the algorithm's ability to embed ethical behavior into an agent's actions. In this game, an agent (L) must navigate to a goal while encountering garbage along its path. The ethical challenge is for the agent to dispose of the garbage correctly such as 'in a bin' rather than taking actions that might be individually beneficial but ethically wrong such as 'throwing the garbage at another agent' (R). Through this experiment, the agent achieved its primary objective of reaching the destination and embraced the desired ethical behavior, for example disposing of garbage in the bin. This dual achievement validates the algorithm's capacity to instill ethical conduct in autonomous agents, effectively aligning their actions with both ethical standards and individual objectives.

2.5 Key Takeaways & Limitations

The key takeaway of this paper is making a significant leap in the field of AI ethics and safety by introducing a novel framework within MORL that ensures autonomous agents can align their actions with ethical standards through a formal guarantee. A standout contribution is the creation of an algorithm designed to automate the construction of learning environments that do more than just encourage ethical behavior. By marrying theoretical insights with practical applications, the paper shows how ethical considerations can be seamlessly integrated into the learning processes of AI agents. This fusion of theory and practice marks a pivotal advancement in AI ethics, laying down a methodological foundation that promises to steer future efforts in developing AI systems that are not only intelligent but also morally aligned. However, the experiment was conducted through a relatively simple Public Civility Game, which raises the question of the framework's effectiveness in more complex scenarios. Additionally, the absence of comparative analysis with existing methods leaves the relative merits and advancements of the proposed algorithm unclear, hindering a full appreciation of its novelty and potential impact in the broader field of AI ethics and safety.

Paper3. Overcoming Blind spots in the Real World: Leveraging Complementary Abilities for Joint Execution [AAAI 2019]

Authors: Ramya Ramakrishnan, Ece Kamar, Besmira Nushi, Debadeepta Dey, Julie Shah, Eric Horvitz

3.1 Key Problem Addressed

The key problem addressed in this paper is formalizing the problem of agent and human blind spot detection in RL to do the safe transfer of control in the real world. Blind spots refer to situations or features that are not adequately represented or absent in simulation environments but are critical in the real world. These blind spots can lead to costly mistakes or unsafe behaviors when the agents are deployed outside of their training environment because the agents are unprepared to handle situations they never encountered during training. The paper focuses on identifying and overcoming these blind spots by leveraging human demonstrations and a novel approach that allows for safe joint execution by determining when to hand off control between the agent and a human operator, thus enabling agents to perform more safely and effectively in real-world applications.

3.2 Summarize

The paper presents an innovative approach to tackling the challenge of blind spots in agents trained in simulated environments when deployed in the real world. These blind spots emerge from discrepancies between simulation and real-world scenarios, potentially leading to unsafe or suboptimal decisions. The core contribution of the paper is a two-step methodology that first identifies critical features missed during simulation-based training through human demonstrations and then learns models to predict both agent and human blind spots for safer real-world execution. This enables strategic hand offs between the agent and human operators, leveraging their complementary abilities. The approach is validated through experiments in two domains, demonstrating its effectiveness in learning to identify and navigate around blind spots, thereby significantly improving safety and performance in real-world applications.

3.3 Strengths & Weaknesses

The paper provides the problem of blind spots due to discrepancies between simulated environments and the real world is soundly formulated. The approach acknowledges the complexity of real-world scenarios that simulations can fail to capture, providing a solid framework for addressing this gap. Moreover, by focusing on a hand-off strategy between AI agents and humans, the paper offers a novel solution that enhances safety and performance in real-world applications. However, it doesn't fully address all open problems and edge cases, notably when high-quality human demonstration data may not be available or when both AI and human operators face unprecedented real-world complexities. Moreover, while it posits its approach as superior to direct simulation-to-reality transfers, a more nuanced comparison with other methods like domain randomization could offer a clearer advantage in terms of efficiency, safety, and performance.

3.4 Experiment & Results

The experiments are designed to validate the hypothesis that identifying and addressing blind spots through human demonstrations and strategic control hand-offs between agents and humans can significantly improve safety and performance in real-world applications. The setup involves two distinct domains: a variation of the game Catcher, which focuses on catching falling fruits with varied attributes, and a simulated highway driving scenario, which requires navigating a vehicle amidst various obstacles, including ambulances not present in the training simulation. In the Catcher domain, the agent is trained to ignore the size of the fruits in the simulation, while the human demonstrations highlight the importance of fruit size, introducing a scenario to test if the system can identify and learn from this blind spot. In the driving simulation, the key blind spot involves ambulances, which are absent in the agent's training environment but are critical for real-world driving safety. The results show that the proposed method effectively identifies these blind spots and significantly outperforms baselines where agents operate solely based on their simulation training or under direct human control. By learning from human demonstrations and correctly predicting when to hand off control between the agent and human, the system demonstrates improved performance and safety in both domains.

3.5 Key Takeaways & Limitations

The paper introduces a viable solution to the critical issue of blind spots in simulation-trained AI agents when transitioning to real world scenarios, through a novel two step methodology that effectively identifies and mitigates these limitations. It emphasizes the significance of human-agent collaboration, leveraging human demonstrations to spot missing features and predict blind spots, underscoring the approach's potential for safer and more effective real-world AI applications. However, the approach of this paper depends on the availability and quality of human demonstration data, which may not always be accessible or consistent. Furthermore, the scalability of this method to handle more complex or diverse real-world scenarios beyond those tested remains uncertain, particularly as the approach presupposes access to certain features that might not be universally available across all applications.

Paper4. Planning, fast and slow: A framework for adaptive real-time safe trajectory planning [ICRA 2018]

Authors: David Fridovich-Keil, Sylvia L. Herbert, Jaime F. Fisac, Sampada Deglurkar, Claire J. Tomlin

4.1 Key Problem Addressed

The key problem addressed is the challenge of trajectory planning for autonomous systems navigating through cluttered environments, where the need for both speed and safety is paramount, and elements such as obstacle locations are unknown beforehand. Existing methods for trajectory planning tend to fall into two broad categories: those that are computationally efficient but offer few if any safety guarantees, and those that can provide stronger safety guarantees but at a high computational cost. The authors introduce a framework that aims to overcome this dichotomy by enabling safe, flexible adaptation of motion plans in real-time, as autonomous systems encounter new obstacles detected by sensors. This is achieved through the concept of "meta-planning," which allows for the safe switching between different planning models with varying speeds and tracking error bounds (TEBs). This approach enables autonomous systems to adapt their motion plans to previously unknown environments, maneuvering differently in the presence of obstacles than they would in free space, all while maintaining strict safety guarantees.

4.2 Summarize

The paper introduces the concept of "meta-planning," which is an innovative enhancement to the Fast and Safe Tracking (FaSTrack) framework that empowers autonomous systems with the ability to dynamically switch between various online planners, each distinguished by unique speeds and TEBs. This approach enables the systems to adapt their motion plans in real-time as they encounter previously unknown obstacles, balancing the need for speed with safety in navigation. The authors are dedicated to detailing a real-time meta-planning algorithm that generates safe, adaptable trajectories by integrating multiple planning models. Through simulations and hardware demonstrations using a small quadrotor, the paper validates the framework's ability to navigate cluttered environments safely and efficiently, maintaining safety margins around obstacles and showcasing the flexibility of the meta-planning approach in unknown environments.

4.3 Strengths & Weaknesses

The paper introduces the meta-planning concept, which enhances the FaSTrack framework. This approach enables real-time adaptation of motion plans based on newly detected obstacles, providing a balance between speed and safety. The empirical advancements are demonstrated through simulations and hardware experiments that validate the framework's effectiveness. Moreover, unlike traditional methods that fall into either efficient but unsafe planning or safe but computationally intensive planning, this approach offers a middle ground with both real-time efficiency and strong safety guarantees. However, the paper does have areas that could be strengthened, such as a more thorough examination of edge cases involving highly dynamic environments, a comparison with a wider range of existing solutions to underscore its advancements, and an expanded discussion on the framework's scalability and computational demands in complex scenarios.

4.4 Experiment & Results

In the paper, both simulation and hardware experiments were conducted to assess the performance of a meta-planning framework with an autonomous quadrotor navigating through dynamically changing, cluttered environments. These experiments, involving simulations of obstacle avoidance and real-world flights with a Crazyflie 2.0 quadrotor, underscored the system's adeptness at adapting its trajectory in real-time, leveraging various planners to optimize for both speed and safety. The framework consistently maintained safe distances from obstacles, affirming its ability to adhere to stringent safety margins through safety controllers derived from offline computations. Moreover, the meta-planning process's efficiency and real-time execution were highlighted, demonstrating the framework's suitability for fast-paced, practical applications. The comprehensive testing and successful outcomes validate the paper's claims about offering a flexible, adaptive, and safe navigation solution, marking a significant advancement in the domain of autonomous system navigation in complex, unpredictable environments.

4.5 Key Takeaways & Limitations

The paper demonstrates not only allows for speed and safety in navigation through cluttered spaces but also extends the FaSTrack framework's utility to scenarios with unpredictable obstacles. The approach's solid safety guarantees, underscored by rigorous offline and online computational methods, are convincingly validated through both simulation and real-world experiments. However, the research also opens up avenues for further exploration, particularly regarding the framework's scalability to larger or more complex environments, its efficacy in highly dynamic environments, and its generalizability across various types of autonomous systems. These areas present opportunities for deepening the understanding and application of meta-planning in the broader field of autonomous navigation.

Paper5. Safe Reinforcement Learning via Shielding [AAAI 2018]

Authors: Mohammed Alshiekh, Roderick Bloem, Rudiger Ehlers, Bettina Konighofer, Scott Niekum, Ufuk Topcu

5.1 Key Problem Addressed

The paper handles the problem of integrating safety considerations into the RL framework in a way that allows for optimal policy learning while enforcing specified safety properties, expressed in temporal logic, without compromising the learning algorithm's ability to converge.

5.2 Summarize

This paper introduces shielded learning frameworks, which significantly enhance the safety of RL processes by integrating a monitoring and correction mechanism, or shield, that ensures actions proposed by the RL agent adhere to predefined safety specifications. This mechanism is meticulously designed to minimally interfere with the agent's learning process, stepping in only to prevent potential safety violations without compromising the learning algorithm's ability to converge to an optimal policy. The concept of shield synthesis from safety specifications allows for the automatic creation of these protective measures based on temporal logic, offering a robust method to maintain safety during both the learning and execution phases. The authors demonstrate the practicality and effectiveness of this approach through varied experimental scenarios, including obstacle navigation, autonomous driving, and water tank management, showcasing not only the framework's ability to safeguard against safety violations but also its potential to enhance learning efficiency. This approach to integrating safety within RL presents a significant advancement in ensuring that learning algorithms operate within safe bounds, offering a solution to the challenge of balancing optimal decision-making with stringent safety requirements.

5.3 Strengths & Weaknesses

The concept of a reactive system shield, which is adeptly addresses the critical issue of maintaining safety during both the learning and execution phases. It marks an improvement of existing solutions by interfacing only when necessary. This approach not only shows minimal interference but also presents automatic synthesis of shields from safety specifications, pushing the boundaries of current capabilities in safe RL. However, the paper falls short in its empirical evaluation by not including comparisons with other safety enforcement methods, limiting the comprehensiveness of its validation. Additionally, it does not thoroughly explore potential edge cases where the shield could fail or behave unpredictably, which is vital for real-world application readiness.

5.4 Experiment & Results

The paper uses various experiment setups to demonstrate the efficacy of its proposed shielding approach in safe RL such as grid world navigation with dynamic obstacles, a self-driving car simulation, water tank level control, and a modified Pacman game. Through these environments, the shielded agents consistently outperformed or matched the learning speed of their unshielded counterparts while strictly adhering to predefined safety specifications. For example, in the grid world experiments, the shield enabled agents to navigate safely and efficiently, even in more complex scenarios with moving obstacles. Similarly, in the self-driving car scenario, the shielded car avoided crashes entirely and learned to navigate faster than without a shield. The water tank control problem showed faster convergence to optimal policies with shielding, emphasizing its role in guiding the learning process away from unsafe states. In the Pacman-like environment, the shielding effectively prevented any collisions with the ghost from the onset of learning, showcasing the shield's ability to integrate safety without hindering the acquisition of task objectives. These results uniformly validate the paper's main claims, underscoring the viability and benefits of integrating safety directly into the RL process through a novel shielding approach that does not compromise learning efficiency or effectiveness.

5.5 Key Takeaways & Limitations

The author's approach is validated across diverse scenarios presenting its adaptability and the benefit of enhancing safety while potentially accelerating the learning of optimal policies. However, the paper acknowledges limitations such as the challenges in abstracting environment dynamics, specifying safety properties, the need for further exploration of scalability, generalization to more complex environments, and a deeper investigation into edge cases and robustness.