A Survey on Deep Reinforcement Learning

Anonymous authors
Paper under double-blind review

1 INTRODUCTION

Deep Reinforcement Learning (DRL) has emerged as a powerful approach for solving complex problems in various domains, ranging from robotics to autonomous driving and game playing Sharifzadeh et al. (2016). The motivation behind DRL research lies in its ability to combine the strengths of deep learning for handling high-dimensional state spaces with reinforcement learning for decision-making and control. This has led to significant advancements in artificial intelligence, making DRL a topic of great importance and relevance to the AI community Wang & Vinel (2020).

The problem addressed in this survey revolves around the challenges and limitations of DRL algorithms, such as overestimation of Q-values, inefficiencies in handling large state spaces, and difficulties in learning from discrete-continuous hybrid action spaces Leibfried et al. (2017); Fu et al. (2019). To overcome these challenges, we propose a comprehensive analysis of various DRL algorithms, focusing on their strengths, weaknesses, and potential improvements. Our research questions aim to identify the key factors that contribute to the success of DRL algorithms and explore novel techniques that can enhance their performance and applicability in different domains.

In the context of related work, we draw upon several DRL algorithms, including inverse reinforcement learning with Deep Q-Networks (DQNs) Sharifzadeh et al. (2016), cross Q-learning algorithms Wang & Vinel (2020), and parameterized action DRL algorithms such as P-DQN and MP-DQN Bester et al. (2019). Additionally, we consider multi-agent DRL algorithms, such as Deep Multi-Agent Parameterized Q-Networks (Deep MAPQN) and Deep Multi-Agent Hierarchical Hybrid Q-Networks (Deep MAHHQN) Fu et al. (2019), which tackle problems with discrete-continuous hybrid action spaces.

The main differences between our work and the existing literature lie in our comprehensive analysis and comparison of various DRL algorithms, as well as our focus on identifying novel techniques and strategies for improving their performance. By examining the strengths and limitations of existing DRL algorithms, we aim to provide a deeper understanding of their underlying mechanisms and contribute to the development of more effective and efficient DRL techniques in the future Okesanjo & Kofia (2017).

2 RELATED WORKS

Deep Reinforcement Learning and Inverse Reinforcement Learning Deep reinforcement learning (DRL) has been widely applied to various problems with large state spaces, such as autonomous driving Sharifzadeh et al. (2016). Inverse reinforcement learning (IRL) is an approach that extracts rewards from problems with large state spaces using deep Q-networks Sharifzadeh et al. (2016). One of the strengths of IRL is its ability to generate collision-free motions and human-like lane change behavior after a few learning rounds Sharifzadeh et al. (2016). However, DRL methods often suffer from the overestimation problem, which is exacerbated by function approximation errors in deep Q-networks Wang & Vinel (2020). To address this issue, a novel cross Q-learning algorithm has been proposed, which maintains a set of parallel models and estimates the Q-value based on a randomly selected network, leading to reduced overestimation bias and variance Wang & Vinel (2020).

Parameterized Actions and Hybrid Action Spaces Parameterized actions in reinforcement learning consist of discrete actions with continuous action-parameters, providing a framework for solving complex domains requiring high-level actions and flexible control Bester et al. (2019). The P-DQN algorithm extends deep Q-networks to learn over such action spaces, but it treats all action-parameters as a single joint input to the Q-network, invalidating its theoretical foundations Bester et al. (2019). To address this issue, a novel method called multi-pass deep Q-networks (MP-DQN) has been proposed, which significantly outperforms P-DQN and other previous algorithms in terms of data efficiency and converged policy performance on various domains Bester et al. (2019). Furthermore, deep multi-agent parameterized Q-networks (Deep MAPQN) and deep multi-agent hierarchical hybrid Q-networks (Deep MAHHQN) have been proposed for multi-agent problems with discrete-continuous hybrid action spaces, showing significant performance improvements over existing independent deep parameterized Q-learning methods Fu et al. (2019).

Addressing Q-value Overestimation and Policy Gradient Methods To address the problem of Q-value overestimation in DRL, an intrinsic penalty signal has been introduced, encouraging reduced Q-value estimates Leibfried et al. (2017). This algorithm outperforms other algorithms like deep and double deep Q-networks in terms of both game-play performance and sample complexity Leibfried et al. (2017). Policy gradient methods are widely used for control in reinforcement learning, particularly for continuous action settings Okesanjo & Kofia (2017). The off-policy stochastic counterpart to deterministic action-value gradients has been studied, as well as an incremental approach for following the policy gradient instead of the natural gradient Okesanjo & Kofia (2017). A recent work has provided the first off-policy policy gradient theorem, developing a new actor-critic algorithm called Actor Critic with Emphatic weightings (ACE) that approximates the simplified gradients provided by the theorem and finds the optimal solution Imani et al. (2018).

Novel Algorithms and Frameworks Several novel algorithms have been proposed to improve the performance of reinforcement learning, such as STOchastic Recursive Momentum for Policy Gradient (STORM-PG) Yuan et al. (2020), diversity actor-critic (DAC) for sample-efficient exploration Han & Sung (2020), and Stackelberg actor-critic algorithms that model the actor and critic interaction as a two-player general-sum game Zheng et al. (2021). Discriminator-Actor-Critic has been proposed to address the implicit bias present in the reward functions used in adversarial imitation learning algorithms and reduce policy-environment interaction sample complexity Kostrikov et al. (2018). Furthermore, a transfer learning framework called Learning to Transfer (L2T) has been proposed to automatically determine what and how to transfer by leveraging previous transfer learning experiences Wei et al. (2017).

In summary, deep reinforcement learning has made significant progress in recent years, with numerous algorithms and techniques proposed to address various challenges, such as overestimation, parameterized actions, hybrid action spaces, and policy gradient methods. These advancements have led to improved performance and efficiency in a wide range of applications, demonstrating the potential of deep reinforcement learning in solving complex real-world problems.

3 BACKGROUNDS

Deep Reinforcement Learning (DRL) is a subfield of machine learning that combines deep learning and reinforcement learning to tackle problems with large state spaces and complex action spaces. The central problem in DRL is learning an optimal policy that maps states to actions, maximizing the cumulative reward in a given environment Sharifzadeh et al. (2016).

3.1 Foundational Theories and Concepts

Reinforcement learning (RL) is a framework for learning sequential decision-making tasks, where an agent interacts with an environment to achieve a goal. The agent's objective is to learn a policy $\pi(a|s)$, which is a probability distribution over actions a given a state s. The agent receives a reward r_t at each time step t and aims to maximize the expected cumulative reward, also known as the return $G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$, where $\gamma \in [0, 1]$ is a discount factor Wang & Vinel (2020).

Q-learning is a popular RL algorithm that estimates the action-value function $Q^\pi(s, a)$, representing the expected return when taking action a in state s and following policy π thereafter. The optimal action-value function $Q^*(s, a)$ is defined as the maximum expected return achievable by any policy. Q-learning updates the action-value function iteratively using the Bellman equation:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right],$$

where α is the learning rate and s' is the next state Fu et al. (2019).

Deep Q-Networks (DQN) extend Q-learning by using a neural network to approximate the action-value function. DQN addresses the instability and divergence issues in traditional Q-learning by introducing experience replay and target networks. Experience replay stores past experiences in a buffer and samples mini-batches for training, breaking the correlation between consecutive samples. Target networks are used to fix the target values during updates, reducing the risk of divergence Leibfried et al. (2017).

3.2 Mathematical Notations and Equations

In DRL, the state and action spaces are often high-dimensional and continuous. The policy $\pi(a|s)$ and value function $V^{\pi}(s)$ are approximated using deep neural networks with parameters θ and ϕ, respectively. The policy gradient theorem provides a way to update the policy parameters by following the gradient of the expected return:

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi \left[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \right],$$

where $J(\theta)$ is the objective function to be maximized Okesanjo & Kofia (2017).

Natural policy gradient methods, such as Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO), improve the convergence and stability of policy gradient methods by incorporating second-order information into the update rule. The natural gradient is defined as $\tilde{\nabla}_\theta J(\theta) = F^{-1}(\theta)\nabla_\theta J(\theta)$, where $F(\theta)$ is the Fisher information matrix, capturing the curvature of the objective function van Heeswijk (2022).

Actor-critic algorithms combine the policy gradient methods with value function approximation to reduce variance in the updates. The critic estimates the value function $V^{\pi}(s)$, while the actor updates the policy parameters using the advantage function $A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$, which measures the relative value of taking action a in state s Imani et al. (2018).

3.3 Extensions and Variants

Several extensions and variants of DRL algorithms have been proposed to address specific challenges, such as handling parameterized action spaces Bester et al. (2019), multi-agent problems with hybrid action spaces Fu et al. (2019), and transfer learning Wei et al. (2017). These algorithms often build upon the foundational concepts and principles of DRL, adapting them to address the unique challenges of their respective problem domains.

References

Craig J. Bester, Steven D. James, and George D. Konidaris. Multi-pass q-networks for deep reinforcement learning with parameterised action spaces. *arXiv preprint arXiv:1905.04388*, 2019. URL http://arxiv.org/abs/1905.04388v1.

Haotian Fu, Hongyao Tang, Jianye Hao, Zihan Lei, Yingfeng Chen, and Changjie Fan. Deep multi-agent reinforcement learning with discrete-continuous hybrid action spaces. *arXiv preprint arXiv:1903.04959*, 2019. URL http://arxiv.org/abs/1903.04959v1.

Seungyul Han and Youngchul Sung. Diversity actor-critic: Sample-aware entropy regularization for sample-efficient exploration. *arXiv preprint arXiv:2006.01419*, 2020. URL http://arxiv.org/abs/2006.01419v2.

Ehsan Imani, Eric Graves, and Martha White. An off-policy policy gradient theorem using emphatic weightings. *arXiv preprint arXiv:1811.09013*, 2018. URL http://arxiv.org/abs/1811.09013v2.

Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. *arXiv preprint arXiv:1809.02925*, 2018. URL http://arxiv.org/abs/1809.02925v2.

Felix Leibfried, Jordi Grau-Moya, and Haitham Bou-Ammar. An information-theoretic optimality principle for deep reinforcement learning. *arXiv preprint arXiv:1708.01867*, 2017. URL http://arxiv.org/abs/1708.01867v5.

Yemi Okesanjo and Victor Kofia. Revisiting stochastic off-policy action-value gradients. *arXiv preprint arXiv:1703.02102*, 2017. URL http://arxiv.org/abs/1703.02102v2.

Sahand Sharifzadeh, Ioannis Chiotellis, Rudolph Triebel, and Daniel Cremers. Learning to drive using inverse reinforcement learning and deep q-networks. *arXiv preprint arXiv:1612.03653*, 2016. URL http://arxiv.org/abs/1612.03653v2.

W. J. A. van Heeswijk. Natural policy gradients in reinforcement learning explained. *arXiv preprint arXiv:2209.01820*, 2022. URL http://arxiv.org/abs/2209.01820v1.

Xing Wang and Alexander Vinel. Cross learning in deep q-networks. *arXiv preprint arXiv:2009.13780*, 2020. URL http://arxiv.org/abs/2009.13780v1.

Ying Wei, Yu Zhang, and Qiang Yang. Learning to transfer. *arXiv preprint arXiv:1708.05629*, 2017. URL http://arxiv.org/abs/1708.05629v1.

Huizhuo Yuan, Xiangru Lian, Ji Liu, and Yuren Zhou. Stochastic recursive momentum for policy gradient methods. *arXiv preprint arXiv:2003.04302*, 2020. URL http://arxiv.org/abs/2003.04302v1.

Liyuan Zheng, Tanner Fiez, Zane Alumbaugh, Benjamin Chasnov, and Lillian J. Ratliff. Stackelberg actor-critic: Game-theoretic reinforcement learning algorithms. *arXiv preprint arXiv:2109.12286*, 2021. URL http://arxiv.org/abs/2109.12286v1.