World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

Reimagining robotic surgical precision: Reinforcement learning in swimmer-v1 environments

Durga Chavali [1], Vinod Kumar Dhiman [2] and Siri Chandana Katari [3]

[1] *Manager, IT Application, Trinity Information Services, Trinity Health, Livonia, Michigan, USA.*
[2] *Vice President, Information Technology, Deenabandhu Chhotu Ram University of Science & Technology, Sonepat, India.*
[3] *Student, Department of Computer Science and Engineering (IoT), Vasireddy Venkatadri Institute of Technology, Nambur, India.*

World Journal of Advanced Research and Reviews, 2024, 22(01), 1739–1744

Publication history: Received on 15 March 2024; revised on 22 April 2024; accepted on 24 April 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.22.1.1251

Abstract

The current study focuses on how reinforcement learning algorithms tackle complex tasks, specifically analyzing the Swimmer-v1 environment with the reassembly of a serpentine robot in robotic surgeries. Herein, the review pays close attention to two algorithms- Proximal Policy Optimization (PPO) and Deep Deterministic Policy Gradients (DDPG)- focusing on exploration strategies in the Swimmer-v1 environment. Of particular importance here is the mentioning of the fact that the scope of exploration includes the use of parameter noise. Findings show that the DDPG learning algorithm faces outstanding difficulties with local maxima convergence. PPO emerged as the first in terms of algorithm category studied despite continuing issues of high variance. The use of a novel method which consists of tempering the range of variation of standard deviation in action noise gives promising results and can be a road to future improvement and exploration. The study provides a critical understanding of the underlying complexities that may lie hidden within the existing reinforcement learning algorithms. It brings up for repair weak points, particularly in the development of exploration capabilities and convergence stabilities.

Keywords; Swimmer-v1 Environment; Proximal Policy Optimization (PPO); Surgical Robotics; Robotic Reassembly; Reinforcement Learning; Algorithmic Adaptability; Policy-based algorithm; Stability Analysis

1. Introduction

Mimicking natural movements, even down to the simplest ones used by ordinary animals in locomotion, has been very hard since the beginning of computers. Not until these past few years have the tech progress accelerated us almost on the verge of a solution, which is one of the most intricate problems. This paper will focus on the cutting-edge technologies, that brought us to this significant turning point, where the possibility of recreating animal movements with a level of faithfulness enabled by computational systems that is within the reach, if not just around the corner. In the past, mimicking natural movements to such an extent has been difficult and demanding, requiring a high proficiency in biomechanics, physics, and computational modeling. With the technological revolution that enables profound computational simulations seemingly within arm's reach, it is a real possibility that every smallest detail of animal motion, even its subtleties, may be imitated.

One of the major orientations of the thesis is the investigation and implementation of the recently developed technologies that enable uniform and oscillatory movement like a snake. Often, the motion appears to be simple, but it is a complex interplay of muscular contractions, frictions at the contact points with the environment, and the ability to adjust the motion in response to changing conditions. This research aims to take advantage of the breakthroughs in

* Corresponding author: Durga Chavali

recent developments to build a computational framework that would create a snake-movement simulator that could be used as a bio-inspired robotics and artificial intelligence tool. The paper aims to resolve the complexity of snake movement by using a technique that explains well the application of emerging technologies in animal locomotion simulation, which is an enormous step towards the point of balancing the biological and the artificial.

1.1. Motivation

There has been an introspective debate in the current setting of whether or not the controversial topic of robots replacing human beings in their jobs is a good or bad decision. Substantial advancements have been achieved in robot-assisted surgery [1], the blueprint to existing snake robotics predominantly focuses on the preliminary structural design, control, and human–robot interfaces, with features that have not been particularly explored in the literature. The employment of medical robots for cost-effective surgeries has accelerated their immense implementation in hospitals and clinics throughout the world. As reported by CMR Surgical (https:(accessed on 8 November 2021), an exemplary healthcare institution based in the UK, the national clinical area had around 70 robots working in it. In particular, the robotic platforms were used for approximately one million interventions in the world by approximately one thousand five hundred robotic platforms. The soaring growth of robotic surgery demand has been illustrated by the robotic operation worldwide number that increased by 178% in 2015, mainly in the US, with a major focus on different specialties.

While these snake robots exhibit superiority at entering small spaces while performing medical surgery where human intervention is impossible, they still require human control to operate. While imagining the situation in which the snake-shaped robots themselves find the most convenient path of action from their current position is where the reinforcement learning (RL) is critical. Deep Reinforcement Learning is another name for it. It uses deep neural networks to do the job of machine learning without the need for human intervention. It learns through trial and error and nurtures novel solutions to difficult problems.

Despite its possible use, RL is still a poorly studied field where rather obvious tasks can be very intimidating. The recent years have seen many developments in computational power, new software frameworks, and vast datasets as a direct result of which progress in RL is also gained. This is shown by and-through milestones which show that machines can be trained to do certain tasks which before that were believed to be the exclusive preserve of humans.

In recent years, neural networks, systems that try to mimic the inner workings of biological brains, have yielded promising results in these domains. Neural networks are harnessed as flexible tools for modeling non-linear functions to maximize certain

reward functions. This reward is typically a numerical representation of the value of a decision, for example, the points you get in a video game. Neural networks work so well because, unlike rule-based programs, they learn from experience; therefore, even if it is not possible for us to formulate a specific set of rules for a problem, these systems can still find an optimal solution to it.

2. Related Work

The development of surgical endoscopes by Computer Motion, Inc (Goleta, CA, USA) began in the 1990s and was followed by the introduction of the first robotic system, AESOP, by this firm. Afterwards, the developed prototype of ZEUS inherited AESOP's multifunctional qualities including dictation and tactile feedback as described in [2]. Keeping up with the latest technological movements, the iterative development of robotic systems mirrors advancing control schemes with various multichannel ways for conducting biopsies [3], reconstruction, and ablation. Robotic controllers have been successfully controlled or operated using joystick interfaces. Steering, tip angulation, aeration, water suction, rinsing, docking, shaft rotation, and contraction are some of the functions enabled with the use of Ruiter et. al. [4]. The Aer-O-Scope eliminates single-use colonoscopes by utilizing a propulsion-based technique for cannulation and examination of the colon. The capacity of balloons for CO2 insufflation enables the generation of the pneumatic actuation force. Another interesting robotic system, namely IREP, provides bimanual manipulation and operation of its robotic arms when combined with the lights and the visual sensory systems. On the other hand, the limiting factors such as high tensile strength, restricted access to curved pathways, and a single-port entrance are confined.

Over the past few years, RL techniques have been gaining a lot of attention, which signifies their importance. The trend started with the publication of Deep Mind's landmark paper titled " Deep Reinforcement Learning from Pixels " in 2013 [1] not only earning it huge media attention but also triggering a major boost of scientific investigations. Algorithms created by DeepMind had been developed further with different applications by various other researchers. Significant

advancements brought to the original Deep Q Network (DQN) include Double Q-learning, Prioritized Experience Replay, and the Dueling Network Architectures for Deep Reinforcement Learning which are thoroughly studied in this thesis. This exploration within those algorithms consisted mainly of chance transition states caused at some predefined time steps.

Later, the latter brought in the era of deep methods for continuous action spaces with the introduction of the DDPG method in the paper 'Continuous Control with Deep Reinforcement Learning' [5]. Exploration in DDPG was at first achieved by adding noise to predicted actions either correlated with an Ornstein-Uhlenbeck process or in the uncorrelated variant of uniform sampling. Plappert et al. proposed a new technique exploiting noise through the direct manipulation of network parameters rather than action space [6]. Nevertheless, this method has clear benefits for off-policy algorithms such as DDPG, compared to on-policy, since this type of learning directly exploits the policy which is sensitive to any type of noise added to it. However, the method for the on-policy algorithm was introduced though it was found to be negligible when it comes to the continuous control settings under study by this thesis.

Stemming from the DDPG success, the research community introduced new types of stochastic algorithms that managed to achieve breakthrough results in a variety of continuous control tasks, like Swimmer-v1. Major cases are Trust Region Policy Optimization [7] and Proximal Policy Optimization Algorithms [8]. These algorithms have stochastic actions designed into them which make them highly different from the deterministic ones.

3. Implementation

In this section, we will implement two different algorithms, all designed to solve the RL problem. All of the algorithms in this chapter have a different approach to finding an optimal solution. Different extensions will be applied to the algorithms suggested in other papers. In the end, we will also present a new extension to one of these algorithms.

3.1. Deep Deterministic Policy Gradient

Let's now explore the very first algorithm which was presented. It was a paper called "Continuous Control with Deep Reinforcement Learning" written by Lillicrap et al. This network was known as Deep Deterministic Policy Gradient (DDPG) and it would be able to adapt and extend the key ideas behind Deep Q Networks (DQN) to suit scenarios with continuous action domains. Discerned as an actor-critic, the model-free or deterministic algorithm, the DDPG operates successfully across continuous action spaces, based on a deterministic policy gradient framework. The algorithm combines the DPG Theorem according to what is stated in reference [9] with the principles found in DQN success.

Similar to DQN, DDPG also uses an approach that it does not operate directly from observations but only stores the experience of its actions in an Experience Replay Memory. This aims at dissolving the link between various observations and improving the generalized ability of learning. It can be seen that, at every time step, DDPG is producing a mini-batch from the memory to train the neural networks in the algorithm.

Moreover, DDPG involves the idea of target networks, a technique that is found in quite several reinforcement learning algorithms and is used for stabilization of the process of training. Alternatively, DDPG does not utilize the exact original networks for the target values unlike other models, the two instead have soft updates. It is translated to making use of an exponential filter-like updating function at every time step that contributes to the softer process of adjusting the target networks which in turn increases the performance of the algorithm by creating stable and smooth training.

3.2. Proximal Policy Optimization

The passage discusses the utilization of Proximal Policy Optimization (PPO) algorithms in addressing the snake problem within the framework of reinforcement learning (RL), noting that this algorithm, which was released during the paper's composition, represents a novel approach building upon the previously introduced Trust Region Policy Optimization (TRPO), the specifics of which are not detailed here; in contrast to Deep Q Networks (DQN) and Deep Deterministic Policy Gradients (DDPG), PPO distinguishes itself by running its policy for an extended period before updating parameters, thereby significantly reducing computational costs associated with the algorithm, and despite its less frequent parameter updates, it is asserted that PPO demonstrates comparable learning efficiency, implying that it does not require a substantially increased number of time steps to effectively learn in comparison to the aforementioned algorithms[10].

4. Evaluation

The following section will focus on the comparison and contrast of the two different algorithms, the Deep Deterministic Policy Gradient (DDPG) and Proximal Policy Optimization (PPO), in the Swimmer-V1 environment. The paper focuses on the environment immediately above, as it features two joints with omnispec rotational movement capabilities that are similar to a snake-like robot. The swimmer-v1 environment specification details are laid out in Table V.

Table 1 Specification of the Swimmer-v1 environment

Observation space	**Continuous**
Num Observations	8
Action Space	Continuous
Action Dimensions	2

The main goal is to move the robot as far right as possible by providing a small positive or negative reward at each time step, depending on the agent's position relative to its starting point and how smoothly it moves from one spot to another. The experiment seeks to identify the algorithm that produces the most desirable results for a snake-like robot; additionally, the investigation presents a new way of introducing noise into ppo. The same DDPG hyperparameters were used as in past experiments. * DDPG with and without action noise were compared. PPO was first evaluated without action noise, followed by a separate analysis with noise. All experiments ran for 1500 episodes. DDPG took about 10 seconds per episode, while PPO took just a few seconds per episode. * Table VI contains the hyperparameters for all PPO experiments

Table 2 Hyperparameter configuration for PPO

Hyperparameter	**Value**
Layer 1 size	100
Layer 2 size	100
Trajectory size	5
Learning Rate Actor	0.00025
Learning Rate Critic	0.0025
γ	1
λ	0.98
ϵ	0.2

5. Results

The results arising from our diverse algorithms are depicted in Fig.1. Also, local maxima around 120 are a problem for DDPG algorithms as the human agent often faces challenges to work on more complex problems. However, the differences between various runs are minimal for DDPG, the average score being around 120 episodes, but introduction of parameter noise to the model resulted in a marked drop in the performance, and the desired outcome as stated in the original paper could not be achieved for this environment by using the recommended approach.

While PPO emerges with better results, staying at a score of around 200 after 1500 episodes, on the contrary. On the other hand, the PPO has a copious variation in its outcomes. The graphs given below include the best trials for all algorithms only.

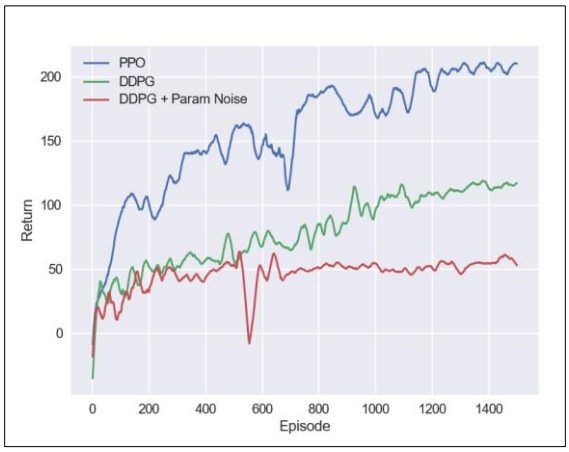

Figure 1 Comparison of the algorithms on the Swimmer-v1 environment

In order to evaluate more our methods, we added a noise to the PPO and compared the results in Figure 11. Like in Figure 10 is the graph of the vanilla PPO that serves as a reference for the other graphs. Retaining a constant standard deviation appears more relaxed because it gets a score of 200 after 400 episodes. Interestingly, a random sampling from a uniform distribution before each trajectory does not actually make any noticeable difference from the original PPO. Importantly, performing a sampling standard deviation from a normal distribution results in a remarkable effectiveness, surpassing 300 after 1500 episodes. This marks a big improvement from the old PPO algorithm and also the others noise introduction methods, and the good results right in the beginning possibly come from a better exploration in the early steps of the learning process.

6. Conclusions and Future Work

The thesis is centered around the comparison of different algorithms aiming to solve the Swimmer-v1 environment's challenges. A major drawback of applying DDPG, it was shown, is the tendency of the policy to stagnate in local maxima rather than searching for a global optimum. Policy gradient-based methods, in particular PPO, ranked the highest among the algorithms tested in the study, which suggests that PPO or policy gradient-based methods, in general, have the best fit for complex, high-dimensional tasks such as the ones we addressed.

The incorporation of a new advance in PPO resulted in a major progress in the topic which we experimented on yet there remain some areas for improvements to make the algorithm efficient and robust. While the different algorithms used in the thesis succeeded in attaining an average return of 96 over 100 consecutive trials; they fell short of the target return of 360 in solving the swimmer-v1 environment. Yet, our most effective way came the closest, which had an average return of 300.

Possible modifications for future improvement of performance also should be done regarding high variability in PPO, but also achieving a more balanced algorithm. Stochastic dynamics of PPO needs another way of exploration with the prospects to wrap the contrasting features of PPO in a deterministic framework, similar to DDPG.

Standard deviation variation exploration provided positive outcome though refinement is needed. Future works may take the form of more sophisticated strategies, such as adapting a trust region which will regulate exploration in terms of net-work-rate of change. Furthermore, one could consider instead varying both the standard deviation and the entire distribution instead of focusing on the standard deviation only to transform it into a uniformization from normal probabilistic distribution to uniform distribution in line with the exploration rate aim of tasks. These considerations include exciting opportunities to develop high-efficiency and stability algorithms implying to tackle complex tasks like Swimmer-v1.

References

[1] Seetohul, J.; Shafiee, M. Snake Robots for Surgical Applications: A Review. Robotics 2022, 11, 57. https://doi.org/10.3390/robotics11030057

[2] Baura, G.D. Medical Device Technologies; Academic Press: Oxford, UK, 2012.

[3] Berthet-Rayne, P.; Gras, G.; Leibrandt, K.; Wisanuvej, P.; Schmitz, A.; Seneci, C.A.; Yang, G.-Z. The i2Snake Robotic Platform for Endoscopic Surgery. Ann. Biomed. Eng. 2018, 46, 1663–1675.

[4] Ruiter, J.; Rozeboom, E.; van der Voort, M.; Bonnema, M.; Broeders, I. Design, and evaluation of robotic steering of a flexible endoscope. In Proceedings of the 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), Rome, Italy, 24–27 June 2012; IEEE: New York, NY, USA, 2012; pp. 761–767.

[5] Xu, K.; Goldman, R.E.; Jienan, D.; Allen, P.K.; Fowler, D.L.; Simaan, N. System design of an insertable robotic effector platform for single port access (SPA) surgery. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, St Louis, MO, USA, 10–15 October 2009; pp. 5546–5552.

[6] M. Plappert, R. Houthooft, P. Dhariwal, S. Sidor, R. Y. Chen,

[7] X. Chen, T. Asfour, P. Abbeel, and M. Andrychowicz, “Parameter Space Noise for Exploration,” https://arxiv.org/abs/1706.01905, 2017, retrieved on 2017.09.26.

[8] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, “Trust Region Policy Optimization,” https://arxiv.org/abs/1502. 05477, 2015, retrieved on 2017.09.26.

[9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” https://arxiv.org/ abs/1707.06347, 2017, retrieved on 2017.10.07.

[10] A. Meena, G. M. V. Reddy and D. P. Chavali, "Accelerated CNN Training with Genetic Algorithm," 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India, 2024, pp. 1-6, doi: 10.1109/IATMSI60426.2024.10502992.