OODA-RL: A REINFORCEMENT LEARNING FRAMEWORK FOR ARTIFICIAL GENERAL INTELLIGENCE TO SOLVE OPEN WORLD NOVELTY

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

14-12-2021 / 28-12-2021

CITATION

Yadav, Pamul; Kim, Taewoo; Suk, Ho; Lee, Junyong; Jeong, Hyeonseong; Kim, Shiho (2021): OODA-RL: A REINFORCEMENT LEARNING FRAMEWORK FOR ARTIFICIAL GENERAL INTELLIGENCE TO SOLVE OPEN WORLD NOVELTY. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.17197913.v2>

DOI

[10.36227/techrxiv.17197913.v2](https://doi.org/10.36227/techrxiv.17197913.v2)

OODA-RL: A REINFORCEMENT LEARNING FRAMEWORK FOR ARTIFICIAL GENERAL INTELLIGENCE TO SOLVE OPEN WORLD NOVELTY

Pamul Yadav

School of Integrated Technology
Yonsei University, South Korea
pamul@yonsei.ac.kr

Taewoo Kim

School of Integrated Technology
Yonsei University, South Korea
boratw@yonsei.ac.kr

Ho Suk

School of Integrated Technology
Yonsei University, South Korea
sukho93@yonsei.ac.kr

Junyong Lee

School of Integrated Technology
Yonsei University, South Korea
jjunilee@yonsei.ac.kr

Hyeonseong Jeong

School of Integrated Technology
Yonsei University, South Korea
jhshseong@yonsei.ac.kr

Shiho Kim

School of Integrated Technology
Yonsei University, South Korea
shiho@yonsei.ac.kr

December 23, 2021

ABSTRACT

Faster adaptability to open-world novelties by intelligent agents is a necessary factor in achieving the goal of creating Artificial General Intelligence (AGI). Current RL framework does not consider the unseen changes (novelties) in the environment. Therefore, in this paper, we have proposed OODA-RL, a Reinforcement Learning based framework that can be used to develop robust RL algorithms capable of handling both the known environments as well as adaptation to the unseen environments. OODA-RL expands the definition of internal composition of the agent as compared to the abstract definition in the classical RL framework, allowing the RL researchers to incorporate novelty adaptation techniques as an add-on feature to the existing SoTA as well as yet-to-be-developed RL algorithms.

1 Introduction

Developing intelligent agents that can interact with and learn from the surrounding environment is one of the primary goals of Artificial Intelligence (AI) research. Among the various sub-fields of AI, Reinforcement Learning (RL) has emerged as the proponent for learning from interaction with the environment via a trial-and-error mechanism [11]. Although, much of the RL theory was developed between late 20th to early 21st century [12, 15, 21, 23, 24], it was mostly confined to solving problems having smaller state space due to the algorithms' lower sample efficiency. Large scale real-world applications of RL were motivated by the introduction of Deep Reinforcement Learning (DRL) [16]. Since then, it has been widely used on domains ranging from video games [17] to robotics [8]. Despite the impressive learning ability of the modern RL algorithms [9, 16, 19, 20], the basis of the RL remains the classical RL framework as shown in Fig. 1. This framework assumes that the agent interacts with the environment by receiving some observations (or states) (o_t) from it, and then it performs an action (a_t) to receive a reward (r_{t+1}) based on the change in the states, where t denotes the time-step. This framework works well for a closed agent-environment interface but fails to incorporate novelties that might occur in the environment without being able to interact with it. Novelty in an open world [13] has been defined by DARPA as previously unseen changes in the environment in their SAIL-ON project description [5], which can be at times, quite adversarial in nature for e.g., in a warfare setting. Therefore, it is crucial to develop the agents which can adapt to the novel changes. Coincidentally, there is an abstract concept called Observation-Oriented-Decision-Action (OODA) loop which was proposed by John R. Boyd to tackle the military warfare problem [2]. It describes an agent to be able to collect information from the environment and orient itself according to the newly gathered information by updating its knowledge and preparing newer strategy in a cyclic manner. A similar approach in the field of Machine

Learning (ML) is known as Active Learning (AL). It involves using an expert system (commonly a Human expert) which supervises the agent in making a decision based on the query [14]. However, it is infeasible to deploy a real-time human expert to resolve every query of the agent in the open world. Inspired by the ideas of OODA and AL, in this paper, we propose an Active Learning-based RL framework which can be easily adapted to develop RL algorithms to solve open world novelty.

Figure 1: The Classical RL Framework.

2 OODA-RL Framework

The proposed OODA-RL framework, as shown in Fig. 2, is an Active Reinforcement Learning based framework that can be used to develop robust RL algorithms capable of handling both the known environments as well as adaptation to the unseen environments. It is designed to incorporate the notion of open world novelty while developing the RL algorithm for solving a real-world problem. It retains the same idea of Environment from the classical RL framework which gives out observations and action-based rewards to the agent. The abstraction of the Agent has been elaborated by defining its internal composition into four stages. Each of the stages are described below in detail. Active Reinforcement Learning based framework that can be used to develop robust RL algorithms capable of handling both the known environments as well as adaptation to the unseen environments.

2.1 Observe Stage

The agent receives raw unstructured data from the environment about the surrounding in the observe stage. Artificial agents can be equipped with a plethora of sensors to collect a diverse type of data such as image, video, audio etc. depending on the problem setting. After collecting the data, it transmits the data to the Orient stage which uses it to develop its own representation of a simulated environment. In parallel, another task of this stage is to process it into meaningful information through representation learning [1]. Processed state representation is feed-forwarded to the Decide stage which in turn decides whether it should pass to Act stage or be provided to Orient stage in case of unseen situation. It is necessary to read in the correct data from the external environment to create appropriate representation. Therefore, state of the art (SoTA) deep learning techniques such as Convolutional Neural Networks [4, 10], Transformers [22], etc. can be used to perform such representation learning.

2.2 Decide Stage

This stage primarily plays the role of a decision maker, where it decides whether the learned state representations are classified as either a seen scenario or an unseen scenario. In case of seen scenario, Decide stage signals the Act stage to use the learned policy for taking an optimal action; whereas, in case of unseen scenario, it sends a query to the Orient stage for finding an appropriate policy to try and adapt to this unseen novelty. Once the adaptation is done, the policy is updated in the Act stage to incorporate the novelty changes as well as updates the newly learned representation in the buffer of Decide Stage.

Figure 2: Proposed OODA-RL Framework for Solving Open World Novelty.

2.3 Orient Stage

This stage plays two primary roles in the framework. Firstly, it creates a continual learning simulated environment based on the raw observation it receives from the Observe Stage. Secondly, it acts as the expert system in the AL loop between Decide and Orient Stage. When the Decide stage sends a query to the Orient stage, depending on the knowledge base, it either explores in the simulated environment or exploits the existing knowledge to find the appropriate policy for novelty adaptation. In an RL setting, the problem of novelty adaptation can be dubbed as an unseen task for the agent. In the past few years, various techniques have started gaining popularity such as Meta-Reinforcement Learning, which allows quick adaptation to the novel changes in the scenario without learning from scratch by utilizing the past experiences [7, 18]. Also, skill is an emerging concept in RL for solving unseen tasks in an extrinsic reward-free setting [3, 6]. Orient stage is also an external reward-free domain which does not interact with the outside world. Thus, the application of Meta-learning and Skill learning can provide suitable bases for developing mechanisms to solve the novelty adaptation problem efficiently.

2.4 Act Stage

A common mathematical representation of the functionality responsible for choosing the action based on the input information in an RL setting is called a policy. In the proposed framework, a policy can be improved (or optimized) by using various SoTA RL algorithms such as DQN [16], PPO [19], SAC [9], etc. to choose the best action based on the state representation. Some of the real-world examples of action are moving a robotic arm, rotating the steering wheel of an autonomous car, etc. Once the action is taken in the real-world, the environment may/may not give an immediate reward depending upon its nature.

3 Conclusion and Future Works

In this paper, we have proposed a primitive OODA-RL framework with the aim of solving the open world novelty. This framework improves upon the classical RL framework by elaborating the agent’s functionality under four stages. It can be utilized for developing robust RL algorithms to tackle the unseen environment changes while maintaining its ability to solve the known world problem. This is a work-in-progress draft which is planned to be updated at frequent intervals until the completion.

Acknowledgements

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-00056, To create AI systems that act appropriately and effectively in novel situations that occur in open worlds)

References

- [1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
- [2] John R Boyd. The essence of winning and losing. *Unpublished lecture notes*, 12(23):123–125, 1996.
- [3] Víctor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giró-i Nieto, and Jordi Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. In *International Conference on Machine Learning*, pages 1317–1327. PMLR, 2020.
- [4] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021.
- [5] DARPA. Science of artificial intelligence and learning for open-world novelty (sail-on). hr001119s0038. 2018.
- [6] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.
- [7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.
- [8] Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE, 2017.
- [9] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.
- [11] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
- [12] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.
- [13] Sangho Lee, Junbeom Park, Ho Suk, Taewoo Kim, Pamul Yadav, and Shiho Kim. An open-world novelty generator for authoring reinforcement learning environment of standardized toolkits. In *International Conference on Multi-disciplinary Trends in Artificial Intelligence*, pages 27–33. Springer, 2021.
- [14] Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6122–6131, 2019.
- [15] David John Cameron Mackay. Introduction to monte carlo methods. In *Learning in graphical models*, pages 175–204. Springer, 1998.
- [16] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- [17] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
- [18] Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. *arXiv preprint arXiv:1803.11347*, 2018.
- [19] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

- [20] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [21] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.
- [22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [23] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.
- [24] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.