Maximum Information Measure Policies in Reinforcement Learning with Deep Energy-Based Model

Conference Paper · April 2021

DOI: 10.1109/ICCIKE51210.2021.9410756

7 authors, including:

Bhopendra Singh
AMITY UNIVERSITY.DUBAI
35 PUBLICATIONS **954** CITATIONS
SEE PROFILE

Regin Rajan
Anna University, Chennai
195 PUBLICATIONS **1,788** CITATIONS
SEE PROFILE

Suman Rajest
Dhaanish Ahmed College
164 PUBLICATIONS **2,775** CITATIONS
SEE PROFILE

Ved P Mishra
Amity University Dubai
80 PUBLICATIONS **767** CITATIONS
SEE PROFILE

2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)
March 17–18, 2021, Amity University Dubai, UAE

Maximum Information Measure Policies in Reinforcement Learning with Deep Energy-Based Model

.K. Sharma
Department of Mathematics
Jaypee University of Engneering and Technology
Raghogarh Dist., MP, India
dilipsharmajiet@gmail.com

Bhopendra Singh
EEE
Amity University
Dubai, UAE
bsingh@amityuniversity.ae

Edwin Herman
Department of Business and Tourism,
Universidad Nacional Santiago Antúnez de Mayolo, Huaraz,
Peru
ehramireza@unasam.edu.pe

R. Regine
Department of Information Technology,
Adhiyamaan College of Engineering
Tamil Nadu, India.
regin12006@yahoo.co.in

S. Suman Rajest
Department of Engineering
Vels Institute of Science, Technology &
Advanced Studies, Tamil Nadu, India.
sumanrajest414@gmail.com

Ved P Mishra
Amity University Dubai, UAE
mishra.ved@gmail.com

2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) | 978-1-6654-2921-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCIKE51210.2021.9410756

***Abstract*—we provided a framework for the acquisition of articulated electricity regulations for consistent states and actions, but it has only been attainable in summarised domains since then. Developers adapt our environment to learning maximum entropy policies, leading to a simple Q-learning service, which communicates the global optimum through a Boltzmann distribution. We could use previously approved amortized Stein perturbation theory logistic regression rather than estimated observations from that distribution form to obtain a stochastic diffusion network. In simulated studies with underwater and walking robots, we confirm that the entire algorithm's cost provides increased exploration or term frequency that allows the transfer of skills between tasks. We also draw a comparison to critical actor methods, which can represent on the accompanying energy-based model conducting approximate inference. Misleading multiplayer uses the recompense power to ensure that the user is further from either the evolutionary algorithms but has now evolved to become a massive task in developing intelligent exploration for deep reinforcement learning. In a misleading game, nearly all cutting-edge research techniques, including those qualify superstition yet, even with self-recompenses, which achieves enhanced outcomes in the sparse re-ward game, often easily collapse into global optimization traps. We are introducing another exploration tactic called Maximum Entropy Expand (MEE) to remedy this shortage (MEE). Based on entropy rewards but the off-actor-critical reinforced learning algorithm, we split the entity adventurer policy into two equal parts, namely, the target rule and the adventure policy. The explorer law is used to interact with the world, and the target rule is used to create trajectories, with the higher precision of the targets to be achieved as the goal of optimization. The optimization goal of the targeted approach is to maximize extrinsic rewards in order to achieve the global result. The ideal experience replay used to remove the catastrophic forgetting issue that leads to the operator's information becoming non-normalized during the off-exploitation period. To prevent the vulnerable, diverging, and generated by the dangerous triad, an on-policy form change is used specifically. Users analyse data likening our strategy with a region technique for deep learning, involving grid world experimentation techniques and deceptively recompense Dota 2 environments. The case illustrates that the MME strategy tends to be productive in escaping the current paper's coercive incentive trap and learning the correct strategic plan.**

Keywords*— *Q-learning; deep reinforcement learning; Boltzmann distribution; Maximum Entropy Expand (MEE);

I. INTRODUCTION

Expert System is one of the most successful or highest machines learning sectors aimed at creating the best outcome to maximize the accumulated benefits gathered in the environment. In many decades, Deep learning - based methods in board games such as Go (AlphaGo/AlphaGo Zero[1,2]), avid gamers similar

978-1-6654-2921-4/21/$31.00 ©2021 IEEE

to Atari[3,] StarCraft II[4], and intelligent method [5,6], and many other variables have had a significant influence on the implementation of deep learning, which has increasingly become a popular way of moving towards general artificial intelligence. However, there are major obstacles in the complex modern world as well when it is of particular importance to conduct effective environmental exploration, strong in the context of incentive or dunord tasks. The sparse motivation the difficulty allows the user to take from the environment a particular sequence of events to obtain the compensation. Even the dishonest incentive problem is more complicated because the specious incentive feature will deceive the agent on discovery di- rection rather than feedback has rarely been given, leading the worker to fall into the ideal local pit. "The issue of "hard-exploration" is the sparse motivation and deceptive reward problems, and the standard reinforcement algorithm does not resolve the problem effectively [7]. However, the "tricky" question probably boils down to the most complicated real-world problems. The unintended local optimum would sometimes induced only. Although we are keen to supply abstract objectives, or with this entire question, certain encouragement mechanisms would not provide greater clarity. To solve scarce recompense problems, some of these standard research methods may split into four broad groups: 1) ordinary defined incentives axioms based on personal knowledge; 2) imitation learners involving specialist incident; 3) Probabilistic discovery method. Initially, inside the limited bandwidth, the manipulative game can reap the benefits of humanity's cultural prejudice or AI to model its reward function that allows its investigator to pursue an inadequate strategy that makes it much harder for the game to reduce the strength of the endogenous game. During the termination of the agent-environment interaction, we conclude that the accrued reward might be positive but raise.

At almost the same level, major difficulties, such as data mismanagement, uncertainty in the training period, remain with challenging exploration problems. Therefore, it is important to have an efficient discovery method that will disassociate exploitation from exploration and test efficiently. Knowledge repeats always used to increase major strategy in many methods. The expertise regarding agent interaction with the environment placed in a hedge and during the recruitment process and the expenichtsce replayed. As a result, the important and worthwhile states will be included in the hedge, guiding the agent to a higher remuneration role. In addition, since it cantered on manipulating the environment large enough to activate it, the operator can no longer investigate the environment. The total entropy of reinforcement learning to optimize the interventions, both strategy's expected entropy and the increased utility, raises the traditional goal RL and boosts the attorney's discovery potential by obtaining different ways. It is a standard technique of undirected manipulation that can escape a covalent bond is formed. Then again, maximum entropy is entangled in a particular plan, like the Soft Action star approach, discovery and exploitation; it can not suit to cope with the dishonest game. We propose an informed exploration method based on the superstition growth strategy to alleviate the challenge of the betrayal incentive dilemma. Now with off-actor-critical prediction model, entropy gains. The employee benefits are the accumulated entropy that would minimize the monetary recompenses by abusing it. We split the actor-critical Algorhythm into provisions: the target rule and the icon's rules, to prevent investigation and aggression intervention. The traverser's policy can evaluate the scheme of cumulative entropy optimization and the directions created by the attack (exploration). It will allow the agent to expand the area where he is aware of the target strategy. To maximize the accumulated external recompense (mistreatment) and attain instructiveness, the target policy can also exploit the explorer's data. To create the optimized experience, replay system that incorporates job recompense (the entropy of help differentiates) with a repeat of knowledge to obtain additional useful knowledge, too strong a concentration seen between exploration policy or the goal strategy the ineffective problem sample and during the training process.

We are implementing a robust on-policy model transition system for the probability of uncertainty induced by the application of off-policy learning, elevated, and non-linear neural network feature equations in switching freely to on and off-policy types. This article presents research to collate our methodology with the policy-based reinforcement learning algorithm and inspection algorithm in such disappointing environments (ICM, RND, SAC). In part 2, related work is given. Entropy in Reinforcement Learning explained in section 3. In section 4, Maximizing for Long-term Entropy is processed. In part 5, Maximum Entropy Policies explained. Section 6 gives the conclusion. The findings show that our strategy successfully prevents and maintains the increasing significance of both the influence of deceptive recompense s in all of these misleading games.

II.RELATED WORK AND BACKGROUND KNOWLEDGE

Betrayal play Deceit play is a scene, which misleads the participant or fellow to use selection dissonance to take a – anti approach. The primary aim of deep learning is through interactions with the environment to increase the compensation gained. Even then, because of short-term incentives in pursuit, the purposely-misleading recompense context makes it very possible for the entire negotiator to fall into the optimum local pit, thereby holding the optimized payoff down the path. Designing the deceiving layout of incentives in the game can make the experience hard while increasing difficulty. On various levels of cognitive beliefs and misleading incentives, the manipulative play can split into three categories: rapacious jaws, sterile traps, and prejudicial jaws. However, the ravening pit carried to trick contributors who trust the renowned incentives so much. Those representatives are committed to optimizing these short strengths and assume that almost all potential benefits will often lead the officer to pick the ideal to bring up for the resolution. The dishonest trap could quickly trick investigators into searching in the opposite prejudice and hardly getting the maximum incentive by placing a tiny number of incentives in the suggested System's reverse direction. Masking traps [8] confuse researchers who think successful approaches are similar to one. In addition, the Monte Carlo query decision tree proffer searching for sources of enhanced returns, but this will cause its agent to disregard the optimization algorithms hiding near the inferior output technique [9]. A established jaws means that agencies prefer to follow the same approach in a similar situation. From the sweeping established presumption, the agent tries to avoid trying to learn through each scene. The generalized imaginary limit, however, therefore is impossible to articulate [10-11]. E.g., in a certain scene, there will also compensated if a surrogate keeps consuming the same amount of gold coins, and if the coins are too high, the compensation will fade into the background. The investigator must, so once again, learn how instead, collect a tonne of precious metals and leave right away. In the deceitful game, the incentive scheme or image conceptually, architecture is logical and does not intentionally mask and/or conceal purposefully. Provide fake intervention. In certain game design, individuals are widespread and extremely likely as the complexity of the match rises in thoughtfully designed recompense systems. The deceptive player analysis allows the agent to select appropriate confirmation prejudice issues and recognize incorrect variables that affect guidance early to accelerate agent preparation. Deep exploration of how sufficient exploration can carried out is a challenging challenge that investigators face today in reinforcement learning. Its goal is to ensure that agents don’t not connected uniquely. To direct the agent to explore the area, most exploration strategies currently focus on trade-off manipulation and depend on dimensional heuristic instructions or purposeful random concept sampling. The greedy method is one of the most winner sic and widely employed techniques. Each pass gives the individual the option of 1-to greedily take the current clinical sorted or the possibility of choosing an individual at random from all policy outcomes. The method is simple to use, has low computational and spatial complexity, and shows outstanding performance on a few simple problems. It is essentially the basis of probabilistic reinforcement learning algorithms. Nevertheless, its methodology does not exploit adequate climate information, nor does it maintain knowledge efficiently. As a result, it is simple to repeat the false behaviour and end up stuttering. Count-based strategies operate on the theory that the more specifically the representatives are aware of the states; the more artefacts are introduced to those states. If the System is inexperienced, both systems will be more perplexed, and operators in grid extension choices will be familiar with the System. A PAC-MDP method shows that the agent can achieve an increasing methodology in multivariate machine learning. Objects examine the sources in the first phase of random samples states experimentally verified. In the second point, agents optimally reconstruct the desired route that has explored. In Montezuma revenge game, he won the highest rank. This technique makes it more difficult to reset freely the environment to any state; it is not feasible in most environments. The fixed goal human brain initialized randomly, as well as the system data acquired by the agent trains the predicted neural net. This method yielded the best results in the Montezuma’s revenge game, with no need to reset to any state. Above all, the method combines exploration and manipulation into a single technique that ignores the manipulative dynamic array influence. As a result, the participant is vulnerable to being deceptive, resulting in the abuse of the accrued incentive. We have introduced a maximum knowledge measure explorer strategy to distinguish exploration from exploitation. The conventional light of the reality strategy to learning the target policy and entropy as an unrevealed framework metric for data generation. The goal rule and the explorer law can also be in effect at the same time.

Deep reinforcement learning (deep RL) has considered a leading route for the checklist to help observable phenomena due to its ability to provide actual working environment sensory input and establish detailed appropriate behaviours through unique neural network representations [3,5]. In what situations is the perfect solution to a stochastic policy? When we examine the relationship between optimization techniques and deterministic reasoning, a probabilistic plan surfaces as the strongest reason, as discussed in previous work. While there are several examples of this model, as an additional consideration, they typically use the cost or incentive function in an original matrix and conclude the optimal conditional distribution over state-conditioned behaviour. The solution shown to optimize an entropy-increased reinforcement lesson objective or match the answer to a matrix factorization-learning task. Intuitively, power for framing.

As an assumption, policies have generated that aim to capture the cheapest available single deterministic behaviour and the entire continuum of low-cost behaviours, maximizing the relevant policy's entropy in particular. Rather than determining the best way to conduct the mission, the resulting procedure tries to learn all how the task conducted. It should now be noticeable, why such policies should be favoured. If we can understand how a given task can conducted, the resulting strategy can act as a useful setup for tweaking to a more particular action. For e.g. first discovering all the directions a machine can push forward, and using this as initialization to learn various running and bounding skills. A better discovery method for discovering the best mode in a multidisciplinary recompense environment; as well as more robust conduct in the face of antagonistic disturbances, where the opportunity to perform the same task in several different situations, Unfortunately, it is difficult to solve such maximum entropy stochastic policy learning problems in a grid setting. A variety of methods have been suggested, including Z-learning, total irreversible entropy RL, approximate induction utilizing communication process [12], -learning, or G-learning, along with deeper structural RL concepts such as PGQ, but these usually function on either simple t How do we apply the process of maximal entropy policy quest to arbitrary policy distributions? In this paper, we borrowed a concept of energy-based models, which in turn reveals an amazing twist between Q-learning, actor-critical computation, and probabilistic inference. In our method, as a (conditional) resource model (EBM), we formulate a stochastic strategy with the energy function related to the 'soft' Q-function obtained when maximizing the maximum target of entropy.

III.PRELIMINARIES

In that same section, the reinforcement-learning problem we are solving will specified, and parameter settings will aim for the maximum entropy system search. We would also pose a few useful identities that we will build upon in our algorithm, which addressed in Section 3. We spend hours talking about Reinforcement Learning, so those who read the report are likely already conscious (RL). Of course, these observations apply to my everyday life and how the liability provided by RL can applied to the world outside ai systems. Reinforcement Learning is a form of Machine Learning dedicated to optimizing behaviours in the form of intangible motivation for those of those unfamiliar. A few years ago, if you would like a summary, humans wrote a series of initial articles on this topic.

It on a first quick look, with both the aim of trying to resolve optimal conduct, the concept of unpredictability can occur unhelpful to an interface. Optimal behaviour is certainly not purposeful. Nonetheless, it points out that spontaneous acts were important in teaching activity. We pick factors that influence RL, and we want our agents to express their worlds. Ostentatious displays are just as powerful a choice to turn other to start travelling the city rather than any prior singularity knowledge. That said, in particular circumstances, such as when selecting behaviour from a wave function, such random action taken, as you can find throughout value-based methods since you would have a vector timetable for trajectory tracking in assessment metrics.

IV.ENTROPY IN REINFORCEMENT LEARNING

In so numerous RL learning rules, such as the strategy or actor-critical families: p (a | s), the measures interpreted as a random variable, conditional mostly on the planet's future. When an agent performs a discrete action, a categorical distribution used, selecting several possible actions. The short comings of many data mining information system can be prevented by real time prevention system [13-16]. In a constant carrier system, a linear function with such a mean and median can used. For such systems, the randomness of the behaviour, a participant takes can be determined by the entropy of the probability distribution, as shown in figure1.

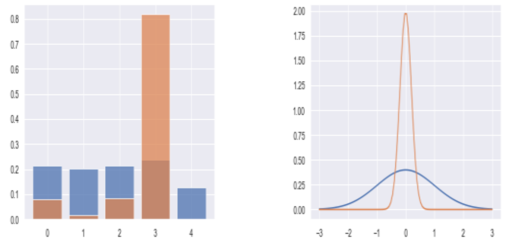

Fig.1 Numerical distributions (left) and Gaussian distribution (right). Orange displays distribution with low-information measure, while blue displays high-entropy distributions.

Information measure is a long-histories word. In mechanics, it typically used to describe the lack of structure within such a system. Then, as a measure of data existing within a conversation, it introduced into the centre of knowledge theory. The theoretical concept of knowledge repurposed in the case of RL. Since RL applicable to learned habits, entropy here explicitly connects to the uncertainty [1] of the behaviours in a given policy taken by an agent.

Sure X be a discrete arbitrary inconsistent with possible values $\{x_1, x_2, \dots x_n\}$ and uncertainty mass function $P(X) = \{p_1, p_2, \dots, p_n\}$, then Shannon's entropy[1] is defined as

$$H(P) = -\sum_{i=1}^{N} p_i \log p_i \qquad (1)$$

Usually, RL aims to maximize the deep number of deferred incentives. This means learning to take a particular course of action to achieve this goal except for other potential action sequences. Such a learning process will inevitably lead to a decrease in the rotation policy's entropy for intervention. This is only fair since the behaviour would necessarily be less spontaneous than the existing bill if we assume purposeful and orchestrated behaviour.

1. Encouraging Entropy

Even so, those who are acquainted with RL writing will know this was not the whole storey. In addition to facilitating a program to focus on acts that lead to high long-term recompense against a probabilities collection, it is often common to add what may be called an "entropy bonus" to the loss function. This benefit allows the agent, instead of less so, to take acts more irregularly.

$$\nabla_{\theta'} log\pi(a_t/s_t;\theta')\big(R_t - V(s_t;\theta_v)\big) + \beta\nabla_{\theta'}H\big(\pi(s_t;\theta')\big) \qquad (2)$$

Renovate equation for A3C. Information measure bonus is **H(π)** term.

Entropy incentives were used by an advisor who may clash too easily with a locally optimal strategy but not necessarily globally optimal without them. Everyone who has worked empirically on RL problems can vouch for the reality of how often an officer could be stuck studying a strategy that only runs into the walls and then only turns in a specific direction or any number of habits, which are obviously. Suboptimal yet low-entropy. Suppose perhaps the global optimum action due to scarce incentives or even other factors is more difficult to grasp. In that case, an individual may also blame for deciding about something simple and less efficient. The entropy bonus used to try to combat this pattern by adding an increasing entropy term to the gradient descent, and it appears to function well in most cases. Indeed, several new states of the art on-policy Deep RL techniques, such as A3C, PPO, and many others, take this stance.

2. Maximizing for long-term entropy

While entropy recompenses commonly utilized, they refer to a more fundamental concept in learning behaviour theory. So far, the entropy bonus offered that year was what has called a single-step bonus. This is because it applies only to the agent's present state in an area and does not consider the possible states where the agent might place. It can think of as a "greedy" entropy optimization. We may draw a parallel to how RL operators learn through incentives. Rather than maximizing the incentive at each point, agents trained to maximize the lengthy amount of potential compensation. We may apply this philosophy to the unpredictability of both assistant's policies or maximize the high number entropy.

A mathematical study from a community of study certainly exists to suggest that the plausible idea is also not to get an entropy bonus at each point of time but to optimize for this great goal. This shows that preparing to get as much potential recompense as practicable is also not optimal with an agent and putting itself in roles where its prospective entropy will also be the greatest.

$$\pi^* = argmax_\pi E_\pi[\sum_{t=0}^{\infty} \beta^t (v_t + \alpha H_t^\pi)] \qquad (3)$$

Development for Maximal Entropy Reinforcement Learning. Optimized architecture π, both decreased, refers to maximum recompenses and information measure, as shown in figure 2.

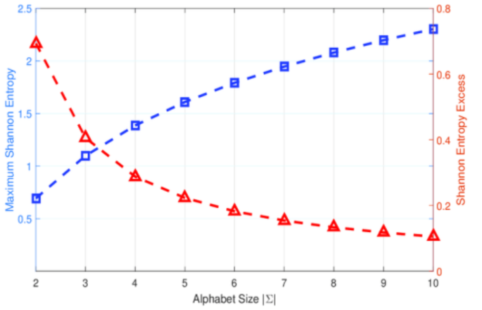

*Fig.2.*Maximum information measure is shown by blue square and entropy excess shown by a red triangle

However, one way of living is that having far more opportunity as feasible is crucial to a value obtained, but is as pro conceivable to the exact number of processes it involves to improve its future actions. The idea that minimizing long-term entropy includes maximizing deep adaptive potential is an incrementally good way of thinking about it. This technique is useful in several ways, leading to improvements in the interpretation of the environment by the agent or changes in the environment over time. As you will see in the graph below, there is some analysis empirically validate this strategy as well.

A tall entropy recompense often results in similar or better performance in these few Mega drive activities that are symmetrical.

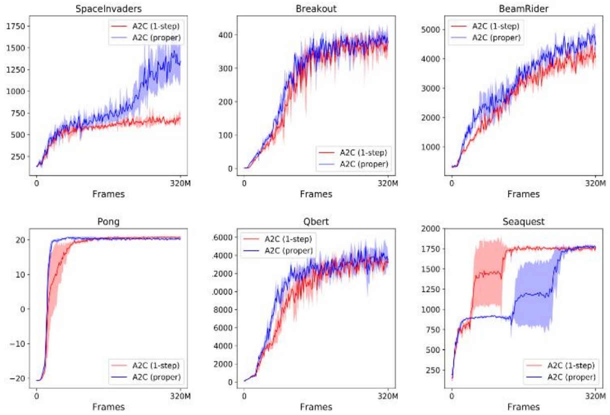

*Fig.3.***Performance by chemometric the one-step (red) entropy bonus to lengthy entropy reduction (blue).

Figure 3 shows that the deep optimization of entropy contributes to as high or higher efficiency as naive optimization of instability in the six-compliance risk.

3. Maximum entropy policies

We also want to continue to argue how this theory of maximum information measure reinforcement learning simply relates to RL and impacts many aspects of life that much more broadly than RL. The fundamental concept of maximum entropy RL is that a balance point among commitment and ability to adapt contributes to optimal behaviour. I guess that matters to life decisions far more than to the actions of virtual intelligence.

Envision the fictional scenario of adjusting to a larger city in a cooler area when you grew up. Maybe you've created a time to dress t-shirts and shorts, where you still came from. It could result in a less satisfying fit in the new city. Your willingness to change your style to suit the past events is directly related to your fashion policy's "high-entropy". You have designed the clothes for comfort in the original town. People will easily be able to adapt to the new city if they have a high-entropy plan. If you have a plan for low-entropy clothes, you could effectively hang on to updates to those posts and suffer as a result. It would be equivalent to not planning to allocate, for example, all the extra cash on T-shirts. The illustration mentioned seems rather stupid, but I believe it reflects a wide range of phenomena that we are experiencing in our lives today. At the social level, let us remember a definition: society's advancement. Taking any trend in science, such as, for example, the Copernican, Darwinian, or your own choice. Scientists trying to maximize the benefits of scientific discovery (fame, reality, effects on society/technology, etc.) gave a chance to either proceed on their post-rate schedule or adjust to the new paradigm. Those with 'high-entropy' research programmes are more likely to respond to a heat universe-based science programme or grow that characteristics of an entity produced from the biological selection. In contrast, those of us with low-entropy reforms are far more prone to stick with their schemes to the detriment of everyone. The prolonged maximum-entropy factor comes early in the careers of these scientists.

V.CONCLUSION

Via an approximate analysis via Stein variation loss function, a method for learning probabilistic energy-based policies was outlined (SVGD). Our method can view as a kind of soft Zed procedure with projected consequences to allow a wide range of multimodal policies. The sampling of busy deadlines as part of SVGD can see as performing an actor's role in an action star dataset. Our experiments show that complex multimodal actions on problems pertaining from toy point mass tasks to virtual moving and swim robotics with complicated speed controller can reliably assessed by one methodology. The training applications of such stochastic methods require enhanced experimentation by pre-training particular stochastic activities in multimodal targets, including compositionality that can effectively turned into challenge behaviours. Although our research explores some possible energy-based policy demands with perfect inference, it would be an exciting way for future studies to further evaluating their ability to represent diverse compositions of behaviour and their capacity for compoundability.

Several early studies have already shown that law learned for various tasks can indeed be compiled to establish new optimal parameters in the form of linearly fixable MDPs [2,3]. Whereas only plain, tractable portrayals have discussed in these previous works, our methodology can use to expand these findings to complicated and multimodal deep neural network models, making them suitable for high-systems such as robotic arms for composable scope. This composability can also use in future projects to construct an enormous multitude of near-optimal capabilities from a catalogue of energy-based plan building blocks.

The target entity is trained and, using collected data, uses an off-policy user reinforcement learning algorithm. Also, the entropy of the aims to achieve is used in unfamiliar states to calculate the target policy's variance. By swapping the external entity with the uncertainty of both the target policy, often in original data, the information for educating the explorers is created so that it the explorer's con- tins to investigate the unknown states. Success of smart concepts and smart knowledge depends on the progressive developments and digital connection [14]. This article utilizes optimal experience repeat and saves the optimal data and is frequently used to enhance the optimization algorithms, and enhance the algorithm's accuracy, aimed at the problem of devastating artificial neural failure just during training. MEE only needs one goal policy or one discovery policy to be enforced. If a systematic and globalized exploration plan is applied, enhancing exploration efficiency by increasingly broad explorer policies with one goal policy is much more beneficial. In fact, there is a lack of clear and impartial evaluation environments due to the equivalent impact of reinforcing machine learning in analytics and the ubiquitous existence of coercive compensation. Building an evaluation framework based on the type of deceptive play and adding various complexity variables is a critical task. A much more robust off-policy type of training is also being used to acquire an additional durable classification model.

REFERENCES

[1] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423.

[2] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hu- Bert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of go without human knowledge, Nature 550 (7676) (2017) 354.

[3] V. Mnih, K. Kavukcuoglu , D. Silver , A .A . Rusu , J. Veness , M.G. Bellemare , A. Graves , M. Riedmiller , A.K. Fidjeland , G. Ostrovski , et al. , Human-level con- trol through deep reinforcement learning, Nature 518 (7540) (2015) 529 .

[4] O. Vinyals, I. Babuschkin, J. Chung, M. Mathieu, M. Jaderberg, W. Czarnecki, A. Dudzik, A. Huang, P. Georgiev, R. Powell, et al., Alphastar: mastering the real- time strategy game starcraft II, 2019,

[5] S. Levine , C. Finn , T. Darrell , P. Abbeel , End-to-end training of deep visuomotor policies, J. Mach. Learn. Res. 17 (2016) 39:1–39:40 .

[6] P. Mirowski , R. Pascanu , F. Viola , H. Soyer , A. Ballard , A. Banino , M. Denil , R. Goroshin , L. Sifre , K. Kavukcuoglu , D. Kumaran , R. Hadsell , Learning to nav- igate in complex environments, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings, 2017 .

[7] M.G. Bellemare , S. Srinivasan , G. Ostrovski , T. Schaul , D. Saxton , R. Munos , Uni- fying count-based exploration and intrinsic motivation, in: Advances in Neu- ral Information Processing Systems 29: Annual Conference on Neural Informa- tion Processing Systems 2016, December 5–10, 2016, Barcelona, Spain, 2016, pp. 1471–1479.

[8] Avijit Mondal, Radha Tamal Goswami, "Enhanced Honeypot cryptographic scheme and privacy preservation for an effective prediction in cloud security, Microprocessors and Microsystems, Volume 81, 2021,103719, https://doi.org/10.1016/j.micpro.2020.103719.

[9] Avijit Mondal, Arnab Kumar Das, Sayan Nath, Radha Tamal Goswami, "Review Study on Different Attack Strategies of Worm in a Network", Webology, Volume 17, Number 2, December, 2020.

[10] D Datta, S Mishra, SS Rajest, (2020) "Quantification of tolerance limits of engineering system using uncertainty modeling for sustainable energy" International Journal of Intelligent Networks, Vol.1, 2020, pp.1-8, https://doi.org/10.1016/j.ijin.2020.05.006

[11] Leo Willyanto Santoso, Bhopendra Singh, S. Suman Rajest, R. Regin, Karrar Hameed Kadhim (2020), "A Genetic Programming Approach to Binary Classification Problem" EAI Endorsed Transactions on Energy, Vol.8, no. 31, pp. 1-8. DOI: 10.4108/eai.13-7-2018.165523

[12] Bhopendra Singh, S. Suman Rajest, K. Praghash, Uppalapati Srilakshmi and R. Regin (2020) Nuclear structure of some even and odd nuclei using shell model calculations. Proceedings of the 2020 2nd International Conference on Sustainable Manufacturing, Materials and Technologies. AIP Conference Proceedings, 2020, https://aip.scitation.org/doi/abs/10.1063/5.0030932

[13] Mishra V.P., Shukla B. (2017) Process Mining in Intrusion Detection-The Need of Current Digital World. In: Singh D., Raman B., Luhach A., Lingras P. (eds) Advanced Informatics for Computing Research. ICAICR 2017. Communications in Computer and Information Science, vol 712. Springer, Singapore. https://doi.org/10.1007/978-981-10-5780-9_22

[14] Mishra, V. P., Shukla, B., & Bansal, A. (2018). Intelligent Intrusion Detection System with Innovative Data Cleaning Algorithm and Efficient Unique User Identification Algorithm. Jour of Adv Research in Dynamical & Control Systems, 10(01), 398-412.

[15] Dsouza, J., Elezabeth, L., Mishra, V. P., & Jain, R. (2019, February). Security in Cyber-Physical Systems. In 2019 Amity International Conference on Artificial Intelligence (AICAI) (pp. 840-844). IEEE.

[16] Mishra, V. P., Dsouza, J., & Elizabeth, L. (2018, August). Analysis and comparison of process mining algorithms with application of process mining in intrusion detection system. In 2018 7th International Conference on Reliability, Infocom Technologies and Optimization(ICRITO) , pp. 613-617. IEEE.