International Journal of Scientific Research & Engineering Trends
Volume 11, Issue 3, May-June -2025, ISSN (Online): 2395-566X

Orthogonal Adversarial Deep Reinforcement Learning For Discrete And Continuous Action Problems

Konka Kishan*, Thuppathi Krishna Sree*, Ramagiri Nissy Jasmine*, Prathikantam Rakshitha*
Department Of CSE (AI & ML) Of ACE Engineering College, India

Abstract- Deep reinforcement learning (DRL) has excelled in video games but remains vulnerable to adversarial attacks. The project unveils Orthogonal Adversarial DRL (OADRL) to improve robustness in both discrete and continuous action spaces. OADRL integrates orthogonal regularization to limit overfitting and adversarial training to enhance resilience. The method assess against standard DRL models, measuring reward stability, adversarial robustness, and generalization. The project presents the OADRL reduces sensitivity to perturbations while maintaining high performance. OADRL improves robustness, ensuring smoother policies and greater resistance to adversarial noise. The insight highlight its potential for real-world applications like robotics and autonomous systems.

Keywords - OADRL, Adversarial Attacks, Robustness, Robotics, Perturbations, Deep Reinforcement Learning.

I. INTRODUCTION

1. Orthogonal Adversarial Deep Reinforcement Learning (OADRL) is a combined approach developed to make deep reinforcement learning (DRL) agents more secure and reliable, especially when faced with adversarial attacks. These attacks are small but carefully designed changes to the agent's input that can confuse it and lead to poor decisions.

2. OADRL tackles this problem by bringing together two key ideas: adversarial training and orthogonal regularization. Adversarial training exposes the agent to such attacks during training, helping it learns how to handle them. At the same time, orthogonal regularization is used to ensure that the features learned by the agent are diverse and less redundant, which makes the learning process more stable and robust. This approach reduces bias in learning by enforcing orthogonality between the regularization term and the DRL objective, which enhances robustness against both gradient-based and, to some extent, black-box attacks. It is compatible with both discrete and continuous action spaces, making it broadly applicable across different reinforcement learning scenarios. Although it introduces additional computational cost, the trade-off results in greater generalization and security, which are critical in deploying DRL agents in safety-sensitive real-world environments.

3. OADRL enhances representation learning by using orthogonal regularization to promote the extraction of diverse, non-redundant features. Simultaneously, the adversarial training component conditions the agent to handle dynamic and uncertain environments by confronting it with perturbations during training. This dual mechanism leads to improved stability, robustness, and adaptability across various benchmark tasks, offering a unified and resilient solution for real-world deep reinforcement learning applications.

II. BACKGROUND OF THE PROJECT

The motivation of Orthogonal Adversarial Deep Reinforcement Learning derives from the increased awareness of the susceptibility of deep neural networks to adversarial attacks, an issue that carries over to DRL agents in rich environments. In early work, the vulnerability of DRL policies was emphasized with the illustration of how small perturbations of inputs could significantly impair performance. At the same time, techniques of adversarial training, initially proposed for supervised learning, were applied to DRL to boost robustness. But a central concern arose: blindly adding adversarial regularization tended to interfere with the agent's overarching goal of receiving maximum rewards.

To overcome this, researchers turned to orthogonalization techniques applied to other fields within machine learning for inspiration, paving the way toward orthogonal adversarial training. This strategy particularly seeks to separate the adversarial defence mechanism from the intrinsic DRL goal, allowing agents to learn resilient policies without compromising performance in both discrete and continuous action spaces. This development extends the existing foundation of adversarial robustness and DRL and

addresses the crucial requirement for reliable and safe AI in practical use.

III. LITERATURE REVIEW

A deep convolutional neural network that produced historic results in the ImageNet Large Scale Visual Recognition Challenge. This paper showed the strength of deep CNNs for image classification, far surpassing earlier approaches. Through the use of innovations such as ReLU activation, dropout, and GPU acceleration, they demonstrated the power of deep neural networks to learn sophisticated image features, triggering a revolution in deep learning and computer vision. Their work revolutionized the field at its core, demonstrating the effectiveness of deep learning for large-scale image recognition tasks.

Simonyan and Zisserman's 2014 paper :
"Deep Residual Learning for Image Recognition," introduced ResNet (Residual Networks), a revolutionary architecture that solved the vanishing gradient problem, making it possible to train much deeper networks. The main innovation of ResNet was using "residual blocks," which involve skip connections, so that the network is able to learn residual mappings rather than direct mappings. This made it possible to build very deep networks, with state-of-the-art performance on image classification problems and had a big influence on the design of following deep learning models.

H. Zhang, H. Chen, C. Xiao, B. Li, D. Boning, and C.-J. Hsieh :
Robust Deep Reinforcement Learning Against Adversarial Perturbations on State Observations," addresses the susceptibility of DRL agents to adversarial attacks. Their approach is to strengthen the robustness of DRL by resisting perturbations added to the agent's state observations. They advance techniques to train DRL agents that are immune to such adversarial attacks and hence enhance their reliability in actual applications where there may be noise or malicious interference. This paper emphasizes the necessity of adversarial robustness for DRL, especially for safety-critical use cases.

Goodfellow, Shlens, and Szegedy's :
Describing and leveraging adversarial examples," formulated the notion of adversarial examples: inputs that deceive neural networks. They demonstrated that even slight, imperceptible modifications to images could lead deep learning models to misclassify them with very high confidence. Their work exposed the weakness of neural networks and initiated the exploration of adversarial robustness, and also advocated the Fast Gradient Sign Method (FGSM) both as a method for producing such examples, and also as an adversarial training method.

Carlini & Wagner :
This project posed a challenge to adversarial example detection. Their work showed successful bypasses of ten detection algorithms. This uncovered an enormous weakness in deep learning security. It is hard to distinguish adversarial inputs from ordinary data. It highlighted the pressing necessity for solid defense and detection methods.

N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami :
They showed that deep neural networks are vulnerable to well-designed perturbations, causing misclassifications, even when these perturbations cannot be perceived by humans. This work put a spotlight on the intrinsic vulnerability of deep learning to security-critical applications, underlining the importance of strong defenses and further study of adversarial vulnerabilities.

I. Ilahi, M. Usama, J. Qadir, M. U. Janjua, A. Al-Fuqaha, D. T. Hoang, and D. Niyato :
It gives a detailed picture of the weaknesses of DRL against adversarial attacks. It describes the types of attacks launched at DRL agents and examines the performance of the different countermeasures. The paper addresses the difficulties of protecting DRL systems, highlighting the importance of strong defense mechanisms to guarantee their reliability and safety under real-world applications. It is an invaluable source to comprehend the security environment of DRL and its possible solutions in order to mitigate adversarial threats.

L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta :
They proposed a paradigm where an adversary is explicitly integrated into the process of training the RL agent, hoping to discover policies that are insensitive to environment perturbations. By training the agent against an adversary learned during the process, they hoped to generate policies that are capable of dealing with hard and unforeseen circumstances, thereby enhancing the robustness of RL agents for deployment in the real world. This work formed the basis of future research on adversarial robustness in DRL.

K. Ohashi, K. Nakanishi, Y. Yasui, and S. Ishii :
They overcome the susceptibility of DRL to manipulation by adversarial agents by training actors to be resilient against situations where value predictions are manipulated. This is intended to make control policies more robust so that they can perform consistently under adverse circumstances. By addressing the reduction of the effect of manipulated value predictions, the study is helpful towards making DRL systems more secure and reliable.

T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra :
Ongoing control with deep reinforcement learning," presented Deep Deterministic Policy Gradient (DDPG), an algorithm that made DRL possible for addressing continuous control problems. The contribution extended DRL beyond discrete action spaces so that agents can learn policies to perform tasks such as robotics control. DDPG combines the deterministic policy gradient method with deep neural networks to present a stable and efficient framework for learning in continuous action spaces. This work significantly contributed to the advancement of DRL and its application to more real-world control problems.

IV. COMPARSION TABLE

S. No	Title	Author's	Methodology Used	Findings from the Reference Paper
1	"ImageNet Classification with Deep Convolutional Neural Networks"	A. Krizhevsky, I. Sutskever, and G.-E. Hinton	Deep Convolutional Neural Networks, ReLU Activation, Dropout, GPU Acceleration	Key techniques like ReLU, dropout, and GPU acceleration are essential for deep learning.
2	"Deep Residual Learning for Image Recognition"	Simonyan and Zisserman's 2014 paper	Residual Blocks, Deep Architectures, Batch Normalization	Residual connections enable the training of much deeper neural networks.
3	"Robust Deep Reinforcement Learning Against Adversarial Perturbations on State Observation s"	H. Zhang, H. Chen, C. Xiao, B. Li, D. Boning, and C.-J. Hsieh	Adversarial Training, State Observation Perturbation.	DRL agents are vulnerable to attacks on their state inputs. Adversarial training can enhance their robustness.
4	"Explaining and harnessing adversarial examples"	Goodfellow, Shlens, and Szegedy's	Fast Gradient Sign Method (FGSM), Neural Network Analysis.	Neural networks are easily fooled by subtle, crafted input changes.
5	"Carlini & Wagner's project"	Carlini & Wagner	Optimization-based Attack, Attack Detection Bypassing	Existing adversarial defences can be reliably bypassed. Adversarial detection is a very hard problem.
6	"The Security of Machine Learning"	N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami	Adversarial Perturbation, Neural Network Misclassification.	Deep neural networks are inherently suspectable to adversarial attacks.
7	"Adversarial Attacks and Defences in Deep Reinforcement Learning"	I. Ilahi, M. Usama, J. Qadir, M. U. Janjua, A. Al-Fuqaha, D. T. Hoang, and D. Niyato	Literature Review, Taxonomy Development	DRL systems require robust defences against various attacks.
8	"Robust Reinforcement Learning via Adversary Training"	L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta	Adversarial Training in RL, Policy Optimization	Training against adversaries makes DRL agents more

				resilient.
9	"Robust Deep Reinforcement Learning Against Adversarial Attacks on Value Predictions"	K. Ohashi, K. Nakanishi, Y. Yasui, and S. Ishii	Value Function Manipulation, Robust Deep Reinforcement Learning	Attacks on value prediction can hinder DRL. Defending against these attacks improves DRL security.
10	"Continuous control with deep reinforcement learning"	T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra	Deep Deterministic Policy Gradient, Actor-Critic Architecture, Experience Replay	Deep RL can be effectively applied to continuous control tasks.

Table 1: Comparison table

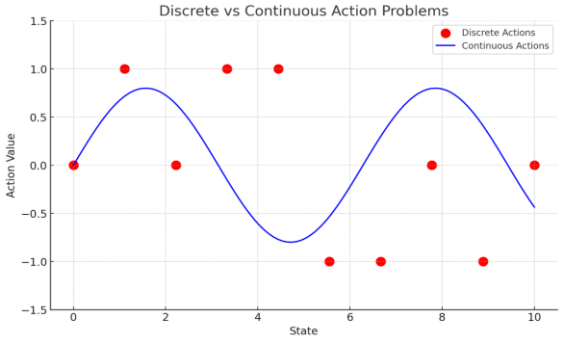

Figure-1: Discrete vs Continuous Action Problems.

V. RESEARCH GAPS

Based on the interrelatedness of those papers, and the state of the field in general, here's an improved examination of the research gaps in current systems:

1. Basic Robustness in Architectures:
• Current architectures (such as those constructed using CNNs and ResNets) are intrinsically susceptible. We don't have architectures specifically constructed to be robust to adversarial attacks from the ground up.

2. Generalizable Adversarial Defences:
• Most defences are specialized and are defeated by new or adaptive attacks.

3.Robustness in Complex, Real-World Scenarios:
•Most current research tends to concentrate on idealized datasets and environments. Real-world situations include noise, uncertainty, and intricate dynamics.

4.Deep Reinforcement Learning (DRL) Robustness Beyond State Observations:
•The majority of DRL robustness studies concentrate on state observation attacks. Attacks on reward functions, action spaces, or even environment dynamics are less studied.

VI. PROPOSED METHOD

Orthogonal Adversarial Deep Reinforcement Learning (OADRL) is a strong, modular system with the aim to enhance the resistance of reinforcement learning (RL) agents to adversarial attacks. OADRL combines orthogonal regularization and adversarial training with standard RL algorithms (e.g., DQN, PPO, TD3). Several orthogonal adversarial generators attack various weaknesses of the agent's internal representation or input. Orthogonal defence mechanisms neutralise these attacks, providing varied threat coverage. Every adversarial path and defense is trained with partially decoupled objectives to promote specialized robustness. Orthogonal regularization prevents the learned representations and policies from being redundant, promoting generalization. OADRL accommodates discrete and continuous action spaces and counters temporal and input-space attacks. By partitioning vulnerabilities into orthogonal subspaces, the system reduces common points of failure. This modular architecture enhances robustness, security, and adaptability to unseen environments. Overall, OADRL offers a promising path for building secure and trustworthy DRL systems in complex real-world tasks.
OADRL Algorithm (3 Steps)

1.Initialization & Observation
Initialize RL algorithm, policy/value networks, adversarial generators Gi, and defence modules Di. Observe initial state s0.

2.Perturbation & Action
Generate orthogonal adversarial perturbations δi=Gi(s),
Aggregate perturbed inputs:s~i=s+δi,
Apply defences: s^i=Di(s~i),
Choose action a∼π(s^i)a({s}_i)a∼π(s^i), interact with environment, store transition.

3. Training & Regularization
Sample from replay buffer, update RL network parameters via:
LRL=E[(r+γamaxQ(s′,a)−Q(s,a))2]
Apply orthogonal regularization to weights WWW:
Lortho=λ‖WᵀW−I‖F2
Jointly update adversarial and defence modules; repeat until convergence.

VII. CONCLUSION AND FUTURE SCOPE

OADRL boosts DRL resilience through orthogonal adversarial generators that address specific vulnerabilities for tailored defences. Modular design promotes generalization and security through the separation of attack dimensions. Scalability in real-world deployments and integration with state-of-the-art DRL methods and verification in realistic environments are essential. Scaling to handle time-based attacks and improving explainability will further boost DRL's reliability. OADRL seeks to develop strong, secure AI systems by training against independent attack vectors and thereby enhancing general stability under adversarial conditions. The research direction of OADRL, encourages modularity and increased security in DRL systems.

REFERENCES

1. .A. Krizhevsky, I. Sutskever, and G.-E. Hinton: Key techniques like ReLU, dropout, and GPU acceleration are essential for deep learning. 2012

2. Simonyan and Zisserman's 2014 paper: Residual connections enable the training of much deeper neural networks

3. .H. Zhang, H. Chen, C. Xiao, B. Li, D. Boning, and C.-J. Hsieh: DRL agents are vulnerable to attacks on their state inputs. Adversarial training can enhance their robustness.2020

4. .Goodfellow, Shlens, and Szegedy's: Neural networks are easily fooled by subtle, crafted input changes. 2014

5. .Carlini & Wagner: Existing adversarial defences can be reliably bypassed. Adversarial detection is a very hard problem. 2017

6. .N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami: Deep neural networks are inherently suseptable to adversarial attacks. 2017

7. .I. Ilahi, M. Usama, J. Qadir, M. U. Janjua, A. Al-Fuqaha, D. T. Hoang, and D. Niyato: DRL systems require robust defenses against various attacks. 2022

8. L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta: Training against adversaries makes DRL agents more resilient. 2017

9. .K. Ohashi, K. Nakanishi, Y. Yasui, and S. Ishii: Attacks on value prediction can hinder DRL.
Defending against these attacks improves DRL security. 2021

10. .T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra: Deep RL can be effectively applied to continuous control tasks. 2015