Know Your Enemy: Identifying and Adapting to Adversarial Attacks in Deep Reinforcement Learning

Extended Abstract

Seán Caulfield Curley
University of Galway
Ireland
s.caulfieldcurley1@nuigalway.ie

Karl Mason
University of Galway
Ireland
karl.mason@nuigalway.ie

Patrick Mannion
University of Galway
Ireland
patrick.mannion@nuigalway.ie

ABSTRACT

It has been shown that an agent can be trained with an adversarial policy which achieves high degrees of success against a state-of-the-art DRL victim despite taking unintuitive actions. This prompts the question: is this adversarial behaviour detectable through the observations of the victim alone? In competitive simulation environments, we find that widely used classification methods such as random forests are only able to achieve a maximum of $\approx 71\%$ test set accuracy when classifying an agent for a single timestep. However, when the classifier inputs are treated as time-series data, test set classification accuracy is increased significantly to $\approx 98\%$. This is true for both classification of episodes as a whole, and for "live" classification at each timestep in an episode. These classifications can then be used to "react" to incoming attacks and increase the overall win rate against Adversarial opponents by approximately 17%. Classification of the victim's own internal activations in response to the adversary is shown to achieve similarly impressive accuracy while also offering advantages like increased transferability to other domains.

KEYWORDS

Adversarial Reinforcement Learning; Deep Reinforcement Learning; Opponent Modelling

ACM Reference Format:
Seán Caulfield Curley, Karl Mason, and Patrick Mannion. 2023. Know Your Enemy: Identifying and Adapting to Adversarial Attacks in Deep Reinforcement Learning: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 2 pages.

1 INTRODUCTION AND BACKGROUND

In competitive environments, the ability to reason about an opponent's past behaviour and using that reasoning to predict what they might do in the future is a crucial technique for success. However, if an opponent's actions seem nonsensical, such as a goalkeeper contorting on the ground, and yet they still win regularly, it is difficult to predict what they will do in the future. This is the conundrum experienced by agents facing adversarial perturbations in opponent policies which are designed to confuse the victim agent to the point of defeat. Natural adversarial observations were first illustrated by Gleave et al. [2] in the 3d simulated physics environments proposed by Bansal et al. [1].

This prompted the question: can agents learn to detect adversarial behaviour before it conquers them? The results we report in this paper demonstrate that widely-used supervised learning models can classify both normal and adversarial behaviour with very high degrees of accuracy. When using a long short-term memory (LSTM) model for classification, the trained model can also be adapted for live classification at each time step in an episode. This LSTM allows the "victim" agent to greatly increase its robustness to adversarial attacks and hence its win rate.

The core idea of this paper is an extension of the work of Gleave et al. [2] who demonstrated that an agent in a multi-agent environment can induce *natural* adversarial observations which significantly affect the performance of its opposing agent. Rather than directly modifying the "victim" agent's observations, the attacker was trained to take actions which induced abnormal activations in the victim, causing it to perform poorly.

To date, the only improvement on Gleave's adversaries was made by giving the attacker access to the actions and observations of the victim [4]. This access allowed the attacker to target specific features of the victim's observations to induce maximum distance from the optimal policy. None of the above papers involved any modelling of the adversarial agent.

Most studies of opponent classification rely on encodings of the game state to model opponent behaviour. For example, Spronck and Teuling [3] used 25 features including number of cities, number of units and population size to model the "preferences" of players in Civilization IV. This is the first paper to classify opponent behaviour solely using the raw observations of the environment and of the other agent. It is also the first to use an agent's own activations as input to a classifier for opponent behaviour.

2 EXPERIMENTAL METHODOLOGY

Experiments in this paper were performed in the zero-sum multi-agent competition environments created by Bansal et al. [1]. We consider two types of classifier inputs:

(1) **Activations:** The outputs from each of the 128 nodes of the victim agent's trained neural network when an observation is fed into the input layer.
(2) **Observations:** A vector with 384 entries, comprising the proprioceptive observations of the adversarial agent's joint angles/velocities/positions, the ball's positioning and other values such as actuator forces.

Datasets were generated by simulating 500 episodes with already trained agents in the *KickAndDefend* environment and outputting the "victim" agent's observation and the resulting activation at

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

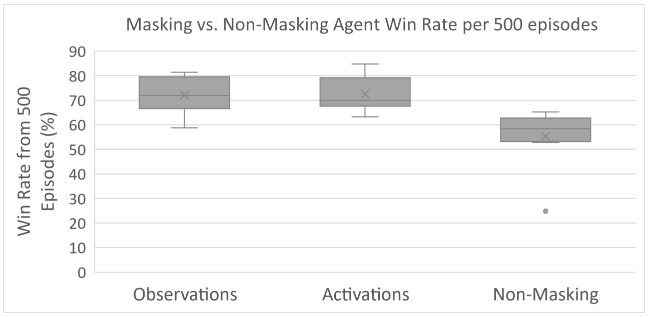

Figure 1: Boxplot of the win rates of both masking and non-masking agents. *Observations* corresponds to the victim reacting by masking based on its classification of its observations while *Activations* represents an agent reacting by masking based on its classification of its activations.

Table 1: Prediction results of in-agent live classification

	True Label	Correct	Incorrect	Unsure
Observ-ations	Adversarial	90.42%	0%	9.57%
	Non-Adv.	98.11%	0%	1.88%
Activ-ations	Adversarial	92.4%	4.8%	2.8%
	Non-Adv.	95.3%	2.6%	1.9%

every timestep. In all cases, the goalkeeper is the opponent agent (which can be Adversarial or Non-Adversarial) and the victim agent is the "kicker".

A number of "traditional" classification algorithms were evaluated; namely random forest, k-nearest neighbours, Gaussian naive Bayes and logistic regression. Grid search cross validation was performed on each model to determine their most effective hyper-parameters and to ensure a fair comparison between models. For time-series classification, activations/observations were grouped by episode and each episode (or time-series) was given a single label corresponding to the opponent agent's behaviour. Both activations and observations were classified using a network comprised of a masking layer, an LSTM layer of 100 units and a single dense output layer. An experiment was also performed where timesteps were inputted one by one with the classifier predicting its output at each step (essentially treating each sequence of timesteps until the last as its own sub-episode).

3 RESULTS

Although the traditional classification techniques tested did learn some of the relationship between activations/observations and opponent behaviour, none performed impressively. The random forest algorithm performed best with approximately 71% accuracy. By contrast, after only 10 epochs of training, the LSTM model achieved $\approx$ 98.0% accuracy on the test set of entire episodes for both activations and observations. However, classifying the entire episode is not the goal of this study. To achieve live classification, a list of all of the timesteps' activations up to that point is passed in and a prediction between 0 (Adversarial) and 1 (Non-Adversarial) is output. A batch of the last 35 predictions is then averaged and the result is classified. If the average of the predictions is less than 0.05, the model predicts that the opponent is acting adversarially. Conversely, if the average exceeds 0.8, Non-Adversarial behaviour is predicted. If the average always remains within the $[0.05, 0.8]$ range, the prediction is deemed "unsure".

To enable in-agent classification, the maximum and minimum values encountered during training for each of the observed features were loaded to be used for normalisation of the live values. Once the episodes began, the classification procedure from above was used to classify the victim's inputs. Each of the three victim agents given by Gleave et al. [2] were evaluated against each of the three Adversarial attackers provided in their paper. Each victim was also evaluated against each of the three Zoo (Non-Adv.) agents provided by Bansal et al. [1]. In all cases, an agent was evaluated against their respective opponent for 500 episodes. The results of this classification are outlined in Table 1.

In real-world applications, it is likely that the user will want the agent to react in some way to a prediction of the opponent type. Therefore, masking (as per Gleave et al. [2]) was implemented. This involves storing the first position of the goalkeeper and using that initial value as a substitute for the actual observed position in order to negate adversarial attacks. Figure 1 illustrates the improvements in win rate as a result of using masking to react to attacks for each masking case for each of the 9 combinations of agents. With masking in the observation classifying case, the average win rate of the 9 victim agents grew from 55.3% to 72.1%. Reacting to one's own activations by masking increased the average win rate across the 9 victim agents to 72.5%. The new win rates show that the classifier can achieve timely, correct predictions which can nullify the effects of Adversarial attacks.

A natural extension of this work is to incorporate a penalty into the reward function of the adversarial agent each time its behaviour can be distinguished from that of a non-adversarial agent in training.

ACKNOWLEDGMENTS

This work is supported by the National University of Galway, Ireland College of Science and Engineering Postgraduate Scholarship.

REFERENCES

[1] Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. 2018. Emergent Complexity via Multi-Agent Competition. *In Proc. ICLR-18* (2018).

[2] A Gleave, M Dennis, N Kant, C Wild, S Levine, and S Russsell. 2020. Adversarial Policies: Attacking Deep Reinforcement Learning. *In Proc. ICLR-20* (2020).

[3] Pieter Spronck and Freek den Teuling. 2010. Player Modeling in Civilization IV. In *Proceedings of the Sixth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (Stanford, California, USA) *(AIIDE'10)*. AAAI Press, 180–185.

[4] Xian Wu, Wenbo Guo, Hua Wei, and Xinyu Xing. 2021. Adversarial Policy Training against Deep Reinforcement Learning. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 1883–1900. https://www.usenix.org/conference/usenixsecurity21/presentation/wu-xian