Grey-box Adversarial Attack on Communication in Multi-agent Reinforcement Learning

Extended Abstract

Xiao Ma
National Key Laboratory for Novel Software Technology,
Department of Computer Science and Technology,
Nanjing University,
Nanjing, China
maxiao@smail.nju.edu.cn

Wu-Jun Li
National Key Laboratory for Novel Software Technology,
Department of Computer Science and Technology,
Nanjing University,
Nanjing, China
liwujun@nju.edu.cn

ABSTRACT

Although research on communication in multi-agent reinforcement learning (MARL) has achieved some progress, the vulnerability of the communication mechanism in MARL caused by adversarial communication messages generated by malicious agents has not been well investigated. Existing works about adversarial communication messages in MARL focus on the black-box scenario where the attacker cannot access any model information about the multi-agent system (MAS). But a more practical setting is the grey-box scenario where the attacker can access the model information about its controlled agent. To the best of our knowledge, there has not been any work investigating grey-box attacks on communication in MARL. In this paper, we propose the first grey-box attack method on communication in MARL, which is called victim-simulation based adversarial attack (VSA). At each timestep, the attacker simulates a victim attacked by other regular agents' communication messages and generates adversarial perturbations on its received communication messages. The aggregation of these perturbations is sent by the attacker to the regular agents through communication messages, which will induce non-optimal actions of the regular agents. Experimental results show that VSA can effectively degrade the performance of the MAS on Predator-Prey. The findings in this paper will make researchers aware of the grey-box attack in MARL.

KEYWORDS

Multi-agent; Reinforcement learning; Communication; Grey-box

ACM Reference Format:
Xiao Ma and Wu-Jun Li. 2023. Grey-box Adversarial Attack on Communication in Multi-agent 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, 3 pages.

1 INTRODUCTION

Research on communication in multi-agent reinforcement learning (MARL) has achieved some progress [1, 2, 4, 5, 7, 9, 11], but the communication mechanism in MARL is still vulnerable [11]. For attackers, their controlled malicious agents might disrupt the collaboration of the multi-agent system (MAS) by sending adversarial communication messages to other agents. Hence, it is necessary for the AI community to study adversarial attacks on communication in MARL. To the best of our knowledge, there exists only one work [10] about adversarial attacks on communication in MARL. This work is a black-box attack in which the attacker cannot access any model of the MAS. However, in real applications, when an attacker has controlled an agent, the model of this controlled agent can be easily accessed by the attacker. This kind of attack is called a grey-box attack and there has not existed any work to study grey-box attacks on communication in MARL.

In this paper, we propose a novel and effective grey-box attack method on communication in MARL, which is called victim-simulation based grey-box adversarial attack (VSA). The main contributions are listed as follows:

- VSA is the first grey-box attack method on communication in MARL. VSA can easily be applied in real scenarios because the grey-box scenario widely exists.
- The attacker simulates the controlled agent as a victim attacked by other agents' communication messages. At each timestep, the attacker generates adversarial perturbations on the received communication messages by disturbing its controlled malicious agent (the simulated victim) to make non-optimal actions. Then the aggregation of these perturbations added to the malicious agent's communication message is sent to other agents.
- Experimental results show that VSA can effectively degrade the performance of the MAS on Predator-Prey [7].

2 METHOD

2.1 Notation

We model a fully cooperative multi-agent task as a decentralized partially observable Markov decision process (Dec-POMDOP) [6] with communication, which can be defined by $G = \langle \mathcal{N}, T, S, O, A, P, R, Z, M, n, \gamma \rangle$. $\mathcal{N} \equiv \{1, 2, ...n\}$ is the finite set of agents. T is the episode time horizon. Here, we take the timestep $t \in \{0, 1, \ldots, T-1\}$ as an illustration. At the timestep t, $s_t \in S$ is the true state of the environment. Each agent $i \in \mathcal{N}$ has its own partially observation $o_i^t \in O$, which is drawn from the true state s_t according to the observation function $Z(s_t, i)$. Agent i sends its communication message $m_i^t \in M$ to other agents $-i$, where M is the communication message space and $-i = \mathcal{N} \backslash \{i\}$. Agent i receives communication messages from other agents $\boldsymbol{m}_{-i}^t = \{m_j^t | j \in -i\}$. Agent i selects an action $a_i^t \in A$ forming a joint action $\boldsymbol{a}^t$ according to the policy $\pi_\theta(a|o_i^t, \boldsymbol{m}_{-i}^t)$, where A is the action space and θ is the

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.

policy parameter. The joint action $\boldsymbol{a}^t$ causes a transition on the environment according to the state transition function $P(s^{t+1}|s^t, \boldsymbol{a}^t)$ and results in a shared reward $r^t = R(s^t, \boldsymbol{a}^t)$. $\gamma \in [0, 1)$ is the discount factor. The existence of attackers aims to degrade the performance of the MAS by sending adversarial messages to other agents and then making other agents take sub-optimal actions. The agent controlled by attackers is called the malicious agent and $\mathcal{B} \subset \mathcal{N}$ is the finite set of malicious agents with size of b. The agent not controlled by attackers is called the regular agent and $\mathcal{R} = \mathcal{N} - \mathcal{B}$ is the finite set of regular agents with size of $n - b$. Malicious agent $i \in \mathcal{B}$ sends the adversarial communication message $\tilde{m}_i^t$ to other agents $-i$ at timestep t. Regular agent $i \in \mathcal{R}$ sends the true communication message m_i^t. We formally denote the communication message sent by agent i as $\hat{m}_i^t$ at timestep t. Hence, we have:

$$\hat{m}_i^t = \begin{cases} \tilde{m}_i^t, & \text{if agent } i \in \mathcal{B} \text{ is malicious,} \\ m_i^t, & \text{if agent } i \in \mathcal{R} \text{ is regular.} \end{cases} \tag{1}$$

2.2 VSA

When attackers generate adversarial communication messages in grey-box scenarios, attackers face two challenges: the occurrence of deadlock and inaccessibility to other agents' data. The deadlock occurs due to the communication messages being sent synchronously and this challenge will be overcome by using outdated messages (shown in Section 2.2.1). As for other agents' data, it is inaccessible naturally in grey-box scenarios, and the victim simulation will overcome this challenge (shown in Section 2.2.2).

2.2.1 Using Outdated Messages. For each agent j, when two timesteps t and $t_1 \leq t$ are close, the communication messages m_j^t and $m_j^{t_1}$ have certain similarities, intuitively. When $\|\hat{\boldsymbol{m}}_{-i}^t - \hat{\boldsymbol{m}}_{-i}^{t_1}\|$ is small, the action taken by agent i will not be changed with a large possibility, which is formulated as follows,

$$\|\pi_\theta\left(a|o_i^t, \hat{\boldsymbol{m}}_{-i}^t\right) - \pi_\theta\left(a|o_i^t, \hat{\boldsymbol{m}}_{-i}^{t_1}\right)\| \leq l_\pi d_i^{t,t_1} \quad \forall a, i, t, \forall t_1 \leq t, \tag{2}$$

where $d_i^{t,t_1} = \|\hat{\boldsymbol{m}}_{-i}^t - \hat{\boldsymbol{m}}_{-i}^{t_1}\|$ and l_π is a constant. Motivated by this, we propose to use received (outdated) messages at the recent timestep $t_1 = t - 1$ to approximate the optimal action at each timestep t, which can overcome the first challenge. Formally, we use $\pi_\theta\left(a|o_i^t, \hat{\boldsymbol{m}}_{-i}^{t-1}\right)$ to approximate $\pi_\theta\left(a|o_i^t, \hat{\boldsymbol{m}}_{-i}^t\right)$. Additionally, the optimal action of agent i at timestep t is denoted as $a_i^{t,*} = \arg\max_a \pi_\theta\left(a|o_i^t, \hat{\boldsymbol{m}}_{-i}^t\right)$ and the approximate optimal action of agent i at timestep t is denoted as $\hat{a}_i^{t,*} = \arg\max_a \pi_\theta\left(a|o_i^t, \hat{\boldsymbol{m}}_{-i}^{t-1}\right)$.

2.2.2 Victim Simulation. Since attackers cannot obtain other uncontrolled agents' data in real applications, we propose to simulate a scenario to generate adversarial communication messages, where the malicious agent under the control of attackers is a simulated victim attacked by its received communication messages. Here, we take the malicious agent $i \in \mathcal{B}$ at the timestep t as an illustration.

In this simulated scenario, the attacker needs to perturb $\hat{\boldsymbol{m}}_{-i}^t$ and then makes the simulated victim i miss the optimal action $a_i^{t,*}$. Due to the existence of the first challenge, we can use $\pi_\theta(a|o_i^t, \hat{\boldsymbol{m}}_{-i}^{t-1})$ to approximate $\pi_\theta(a|o_i^t, \hat{\boldsymbol{m}}_{-i}^t)$, as discussed in Section 2.2.1. Malicious agent i can access its approximate optimal action $\hat{a}_i^{t,*}$. Malicious agent i simulates that it is attacked by the message $\hat{m}_j^{t-1}$ sent by agent j. Hence, the attacker tries to generate a perturbation on $\hat{m}_j^{t-1}$. Here, we introduce a cost function to perturb received messages $\{\hat{m}_j^{t-1}|j \in -i\}$ from agents $-i$ on the malicious agent i, which is shown as follows:

$$J(\theta, o_i^t, \hat{\boldsymbol{m}}_{-i}^{t-1}, \hat{a}_i^{t,*}) = \sum_{a \in A} p(a) \log \pi_\theta(a|o_i^t, \hat{\boldsymbol{m}}_{-i}^{t-1}). \tag{3}$$

Here, $p(a)$ is given by $p(a) = \begin{cases} 1, & \text{if } a = \hat{a}_i^{t,*}, \\ 0, & \text{otherwise.} \end{cases}$

Although our method can be combined with different techniques for crafting adversarial examples, we choose the fast gradient sign method (FGSM) [3] for efficiently generating adversarial examples as an illustration. By minimizing and linearizing the cost function $J(\theta, o_i^t, \hat{\boldsymbol{m}}_{-i}^{t-1}, \hat{a}_i^{t,*})$ around $\{\hat{m}_j^{t-1}|j \in -i\}$, the optimal perturbations under $\ell_\infty-$ norm around $\{\hat{m}_j^{t-1}|j \in -i\}$ are shown as follows:

$$\{\eta_j^{t,i} = \text{sign}\, \frac{\partial J(\theta, o_i^t, \hat{\boldsymbol{m}}_{-i}^{t-1}, \hat{a}_i^{t,*})}{\partial \hat{m}_j^{t-1}} | j \in -i\}. \tag{4}$$

The communication message $\hat{m}_j^{t-1}$ contains the learned information by the agent j, e.g., its historical knowledge and its observations at time-step $t - 1$. Hence, $\eta_j^{t,i}$ as the perturbation of message $\hat{m}_j^{j-1}$ can be used to destroy the learned information by the agent j. In summary, the proposal to simulate victim scenarios can successfully overcome the second challenge.

2.2.3 Generation of Adversarial Communication Messages. At timestep t, malicious agent $i \in \mathcal{B}$ can receive $n - 1$ candidate perturbations $\{\eta_j^{t,i}|j \in -i\}$. Here, we propose to use the mean of the candidate perturbations to aggregate these perturbations, which is defined as $\eta_i^t = \frac{1}{n-1} \sum_{j \in -i} \eta_j^{t,i}$. Then malicious agent i can generate an adversarial communication message $\tilde{m}_i^t = m_i^t + \epsilon \,\text{sign}\, \eta_i^t$ and send $\tilde{m}_i^t$ to others. Here, ϵ is a constant and each element of the message is changed by no more than ϵ.

Table 1: Results of all attack methods. Each attack method is reported with mean ± standard deviation across 5 random runs.

	without Attack	MA	VSA
Win-Rate (%)	100.00 ± 0.00	97.97 ± 1.57	**72.82 ± 3.11**

3 EXPERIMENTS

We evaluate VSA in the predator-prey (PP) task [7], where 2 agents try to reach a stationary prey on a 3×3 grid. Here, we choose CommNet [8] as the communication mechanism in MARL. We use model-based message attack (MA) [10] on communication as the baseline and also use $\ell-$norm as the constraint of the adversarial perturbation in MA for a fair comparison. For all attack methods, $\epsilon = 0.3$ and $b = 1$. The corresponding results of our method and the baseline are shown in Table 1. We can find that VSA can more effectively degrade the performance of MAS than the baseline. The findings of this paper will make researchers aware of grey-box attacks in MARL.

REFERENCES

[1] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. 2019. TarMAC: Targeted Multi-Agent Communication. In *ICML*.
[2] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In *NeurIPS*.
[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR*.
[4] Jiechuan Jiang and Zongqing Lu. 2018. Learning Attentional Communication for Multi-Agent Cooperation. In *NeurIPS*.
[5] Woojun Kim, Jongeui Park, and Youngchul Sung. 2021. Communication in Multi-Agent Reinforcement Learning: Intention Sharing. In *ICLR*.
[6] Frans A. Oliehoek. 2012. Decentralized POMDPs. In *Reinforcement Learning*. Adaptation, Learning, and Optimization, Vol. 12. Springer, 471–503.
[7] Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. 2019. Learning when to Communicate at Scale in Multiagent Cooperative and Competitive Tasks. In *ICLR*.
[8] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2016. Learning Multiagent Communication with Backpropagation. In *NeurIPS*.
[9] Tonghan Wang, Jianhao Wang, Chongyi Zheng, and Chongjie Zhang. 2020. Learning Nearly Decomposable Value Functions Via Communication Minimization. In *ICLR*.
[10] Wanqi Xue, Wei Qiu, Bo An, Zinovi Rabinovich, Svetlana Obraztsova, and Chai Kiat Yeo. 2021. Mis-spoke or mis-lead: Achieving Robustness in Multi-Agent Communicative Reinforcement Learning. *CoRR* abs/2108.03803 (2021).
[11] Sai Qian Zhang, Qi Zhang, and Jieyu Lin. 2020. Succinct and Robust Multi-Agent Communication With Temporal Message Control. In *NeurIPS*.