Jian ZHAO, Youpeng ZHAO, Wenxun WANG, Mingyu YANG, Xunhan HU, Wengang ZHOU, Jianye HAO, Houqiang LI, 2022. Coach-assisted multi-agent reinforcement learning framework for unexpected crashed agents. *Frontiers of Information Technology & Electronic Engineering*, 23(7):1032-1042. https://doi.org/10.1631/FITEE.2100594

Coach-assisted multi-agent reinforcement learning framework for unexpected crashed agents

Key words: Multi-agent system; Reinforcement learning; Unexpected crashed agents

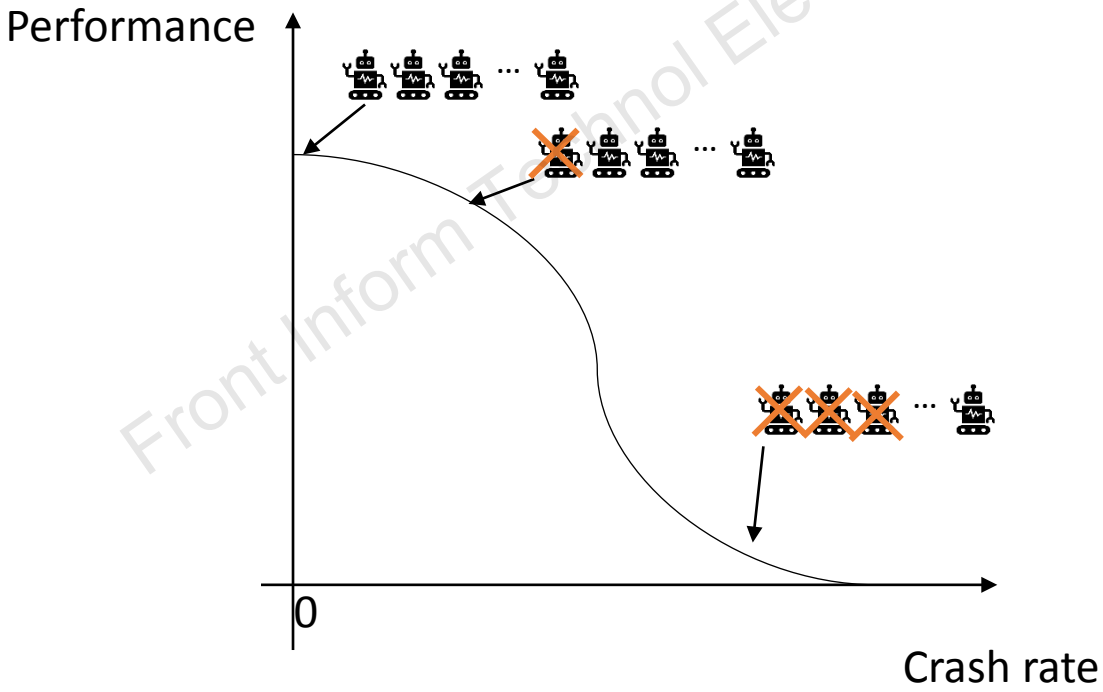
Jian ZHAO
E-mail: zj140@mail.ustc.edu.cn
ORCID: https://orcid.org/0000-0003-4895-990X

Motivation

- Generally, traditional works about multi-agent reinforcement learning (MARL) always train the model in a simulated environment, assuming that agents can work normally all the time. However, this assumption is usually not in line with reality. Because of the inevitable hardware or software failures in practice, one or more agents may unexpectedly “crash” during the coordination process. Once some agents break down during execution, the whole multi-agent system will be greatly affected, resulting in performance degradation.
- For example, when unmanned aerial vehicles fly in a cluster formation, the formation will be disrupted if an agent collapses.

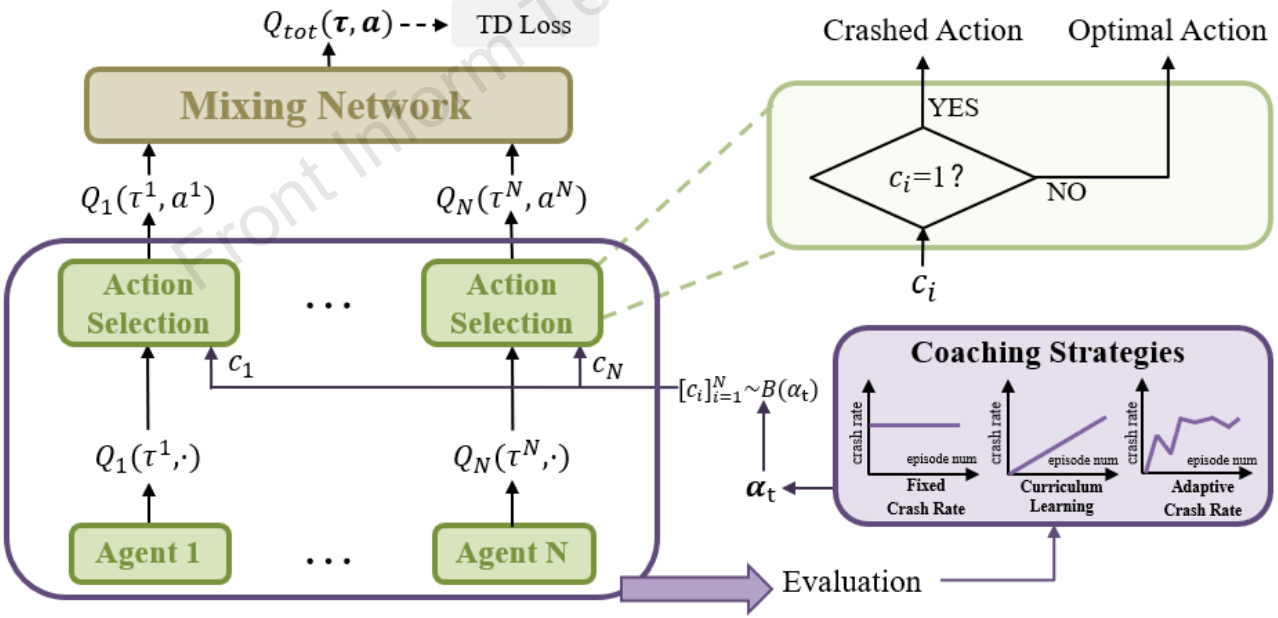
Challenge

- Under the classical centralized training and decentralized execution (CTDE) training schedule, agents cannot communicate with each other and know whether there exist crash agents.
- The number of crashed agents is unsure, so a policy that can handle situations with different crash rates is required.



Method: framework

- Simulate crash scenarios during training.
- Introduce a virtual agent into the system to act as a coach to adjust the crash rate.
  - At the beginning of each episode t, the coach sets up a crash rate α_t. We assume that the probability of being crashed for each agent follows a Bernoulli(α_t) distribution.
  - The coach will adjust the crash rate according to the evaluation results during training.



Method: coaching strategies

- Fixed crash rate
  - The coach sets a fixed crash rate throughout the training. The system has to learn coordination skills under crashed scenarios from scratch.

- Curriculum learning
  - The coach linearly increases the crash rate during training. This approach gradually increases the cooperation difficulty.

- Adaptive crash rate
  - The coach adaptively adjusts the crash rate to correspond to the performance of the agents at the current crash rate. The basic idea is that if the agents can cooperate well and achieve acceptable performance under the current crash situation, the crash rate should be increased; otherwise, the crash rate should be decreased.

Method: resampling strategies

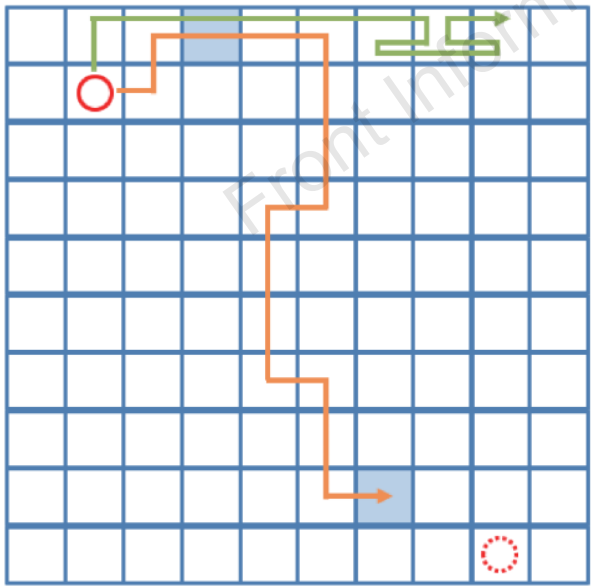
- Whether one agent is crashed is sampled according to a Bernoulli distribution. The independent sampling may cause the proportion of crashed agents to exceed or be smaller than current crash rate α_t.

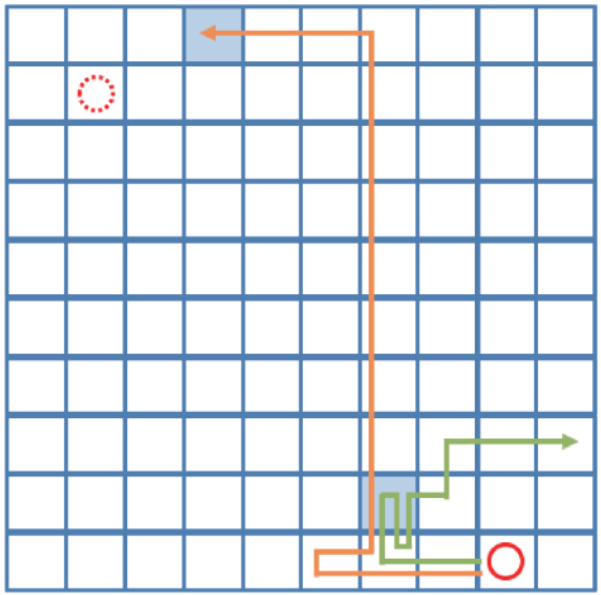
- Employ a resampling strategy to ensure that the number of crashed agents is not larger than the upper bound of $N * \alpha_t$, where N is the total number of agents in the system.

- For the samples with more crashed agents, it may be too difficult for the current model to learn the coordination skills, and thus these samples are discarded and new samples will be generated.

Experiment: grid world

- We conduct an easy experiment to intuitively demonstrate the necessity of considering crash scenarios.
- The two agents (red button) are aimed to reach the two shallowed squares. If one of them is crashed, they cannot move. The green path is the result of the classical MARL algorithm, QMIX, and the orange one is the result of ours. The results show that there exists over reliance on cooperation of systems trained in crash-free environments, leading to a failure when encountering unexpected crashes.





Experiment: SMAC

- The main experiment is conducted on the StarCraft Multi-Agent Challenge (SMAC) and the baseline algorithm is QMIX.
- The main results reveal that our method can achieve good performance under scenarios with different crash rates.

Table 1 The performance of the compared methods in terms of the win rate (including mean and standard deviation) under different crash rates

Method	Win rate					
	3s_vs_5z			3s5z_vs_3s5z		
	Crash rate=0.01	0.05	0.10	Crash rate=0.01	0.05	0.10
Baseline	67.0 ± 15.5	61.9 ± 15.7	56.9 ± 15.3	85.3 ± 11.0	64.8 ± 8.3	43.6 ± 11.3
Fix-0.01	84.8 ± 11.8	74.7 ± 14.5	72.0 ± 13.8	86.9 ± 2.1	63.9 ± 1.7	45.8 ± 2.0
Fix-0.05	84.8 ± 7.5	78.0 ± 12.6	72.7 ± 9.6	86.3 ± 3.1	65.8 ± 2.9	46.9 ± 6.2
Fix-0.10	86.9 ± 8.0	81.1 ± 5.4	74.8 ± 8.2	83.6 ± 3.0	64.8 ± 7.0	48.1 ± 4.2
Curriculum	84.4 ± 6.0	81.1 ± 8.1	74.7 ± 5.8	87.8 ± 2.5	66.1 ± 2.9	48.0 ± 1.6
Adaptive-	82.5 ± 8.5	77.8 ± 8.4	71.6 ± 10.3	85.9 ± 4.5	65.2 ± 3.0	46.1 ± 1.9
Adaptive	$\mathbf{88.6 \pm 3.6}$	$\mathbf{83.3 \pm 6.5}$	$\mathbf{79.2 \pm 6.7}$	$\mathbf{88.0 \pm 3.2}$	$\mathbf{67.0 \pm 2.4}$	$\mathbf{51.7 \pm 2.2}$

Method	Win rate					
	8m_vs_5z			8s_vs_3s5z		
	Crash rate=0.01	0.05	0.10	Crash rate=0.01	0.05	0.10
Baseline	94.1 ± 2.3	82.3 ± 4.5	71.6 ± 2.9	88.6 ± 5.7	75.8 ± 7.0	68.6 ± 4.9
Fix-0.01	86.9 ± 5.4	79.8 ± 5.0	65.6 ± 4.4	87.5 ± 5.8	77.3 ± 6.6	62.0 ± 7.3
Fix-0.05	89.1 ± 2.4	84.2 ± 4.4	68.4 ± 6.4	91.1 ± 6.4	80.0 ± 7.2	66.7 ± 7.2
Fix-0.10	90.0 ± 5.0	83.9 ± 7.1	78.0 ± 4.8	88.9 ± 9.4	79.7 ± 11.3	70.0 ± 9.6
Curriculum	94.1 ± 1.8	82.5 ± 2.8	72.3 ± 2.9	92.0 ± 2.9	79.8 ± 5.4	66.6 ± 5.8
Adaptive-	91.3 ± 4.1	84.8 ± 2.1	78.6 ± 3.1	91.7 ± 3.1	80.0 ± 4.6	69.1 ± 6.0
Adaptive	$\mathbf{94.2 \pm 2.2}$	$\mathbf{89.4 \pm 2.4}$	$\mathbf{81.1 \pm 3.1}$	$\mathbf{93.9 \pm 4.5}$	$\mathbf{84.5 \pm 10.0}$	$\mathbf{71.3 \pm 12.2}$

Fix-i represents the variants of QMIX, indicating that the crashed rate is fixed to i during training. Adaptive- represents the results gained by adopting our adaptive method, but without the re-sampling strategy

Conclusions

- Our coach-assisted MARL framework simulates different random crash rates during training, so that agents can master the skills necessary to deal with crashes.
- The results demonstrated the efficacy and generalization of our method under different crash rates.
- In the future, we will further investigate the case in which crashed agents may take other abnormal actions in addition to random actions and other more efficient coaching strategies.