Discovering Generalizable Multi-agent Coordination Skills from Multi-task Offline Data

Fuxiang Zhang[1, 2*], **Chengxing Jia**[1, 2*], **Yi-Chen Li**[1], **Lei Yuan**[1, 2], **Yang Yu**[1, 2], **Zongzhang Zhang**[1†]
[1]National Key Laboratory for Novel Software Technology, Nanjing University
[2]Polixir Technologies
{zhangfx,jiacx,liyc,yuanl}@lamda.nju.edu.cn
{yuy,zzzhang}@nju.edu.cn

Presented by **Fuxiang Zhang**

Published as a conference paper at ICLR 2023

https://github.com/LAMDA-RL/ODIS

Multi-task Adaptation in MARL

- Cooperative multi-agent reinforcement learning (MARL) faces the challenge of adapting to multiple tasks with **varying agents and targets**.
  - Agent number changes when agents come and go
  - Environment target changes with time

Autonomous warehouse

Environment (10 AGVs)

AGV crashed

New AGV came

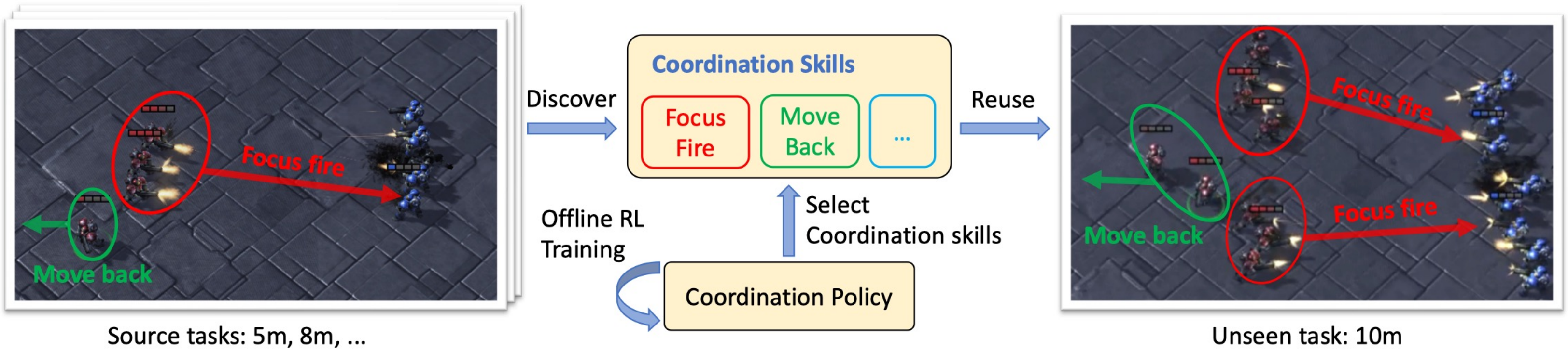
Shelf position changed

MARL methods will fail!

https://images.app.goo.gl/VTiBjtkerC5BGDZa6

Our Motivation

- Exhaustive online interaction is not practical in multi-task settings
  - It can be easier to acquire offline data from a few tasks
  - Can we learn the policy from multi-task offline data and deploy it to other tasks?
- Underlying common structures among tasks can be helpful
  - The coordination skills…



- Our method: An <u>O</u>ffline MARL algorithm to <u>D</u>iscover coord<u>I</u>nation <u>S</u>kills from multi-task data (**ODIS**)

How does ODIS do?

- Deal with varying observation & action shapes for multiple MARL tasks
  - We use flexible Transformer structures to process input of agent sequences
  - For example, agent observations can be decomposed to information of other agents, entities, and the environment, forming embeddings of a sequence.
- Represent and extract coordination skills from offline data
  - The latent coordination skill can recover task-relevant actions

$$L_{\mathrm{s}}(\theta_a, \phi_s) = -\mathbb{E}_{(s,\boldsymbol{\tau},\boldsymbol{a})\sim\mathcal{D}}\left[\sum_{i=1}^{n} \mathbb{E}_{z_i\sim q(\cdot|s,\boldsymbol{a},i)}\left[\log p(a_i \mid \tau_i, z_i)\right] - \beta D_{\mathrm{KL}}\left(q(\cdot \mid s, \boldsymbol{a}, i) \parallel \tilde{p}(\cdot)\right)\right].$$

- Learn a coordination policy to select these skills
  - Enable general policy improvement with traditional centralized training & decentralized execution paradigm in MARL

$$L_{\mathrm{p}}(\theta_v, \phi_o) = \underbrace{L_{\mathrm{TD}}(\theta_v)}_{\text{TD Learning}} + \underbrace{\alpha L_{\mathrm{CQL}}(\theta_v)}_{\text{Conservative term}} + \underbrace{\lambda L_{\mathrm{c}}(\phi_o)}_{\text{Consistent representations}}$$

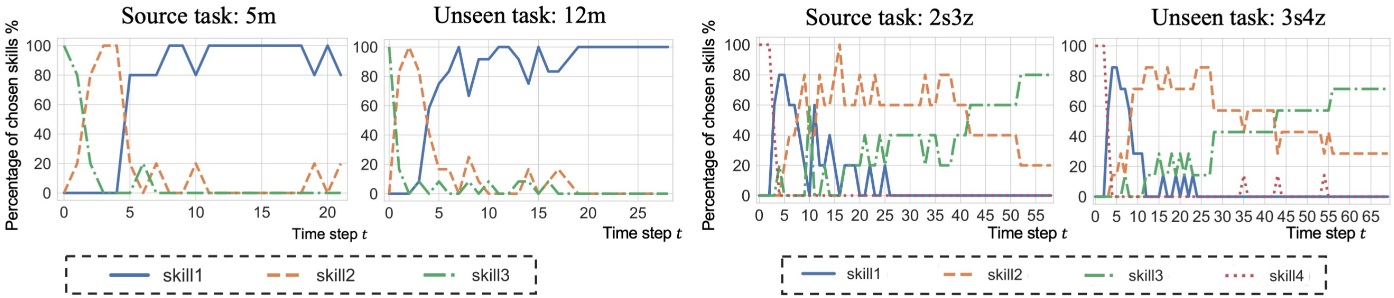
Performance on Multi-task Generalization

- StarCraft II multi-agent Challenge
  - marine-hard task set: containing marine battle offline data on 3 maps
- Baselines:
  - BC-best. Best results from BC-t (standard behavior cloning with Transformer models) and BC-r (add additional return-to-go input like Decision Transformer)
  - UPDeT variants. UPDeT-l use additive VDN mixing network and UPDeT-m uses ODIS's (ours) mixing network

Task	Expert: BC-best	Expert: UPDeT-l	Expert: UPDeT-m	Expert: ODIS (ours)	Medium: BC-best	Medium: UPDeT-l	Medium: UPDeT-m	Medium: ODIS (ours)
Source tasks								
3m	97.7 ± 2.6	71.0 ± 16.6	82.8 ± 16.0	**98.4 ± 2.7**	65.4 ± 14.7	56.6 ± 14.2	51.2 ± 3.4	**85.9 ± 10.5**
5m6m	50.4 ± 2.3	12.1 ± 12.6	17.2 ± 28.0	**53.9 ± 5.1**	21.9 ± 3.4	5.6 ± 4.8	6.3 ± 4.9	**22.7 ± 7.1**
9m10m	**95.3 ± 1.6**	26.6 ± 12.0	3.1 ± 5.4	80.4 ± 8.7	63.8 ± 10.9	34.4 ± 13.9	28.5 ± 10.2	**78.1 ± 3.8**
Unseen Tasks								
4m	92.1 ± 3.5	28.6 ± 21.6	33.0 ± 27.1	**95.3 ± 3.5**	48.8 ± 21.1	21.6 ± 17.2	14.1 ± 5.2	**61.7 ± 17.7**
5m	87.1 ± 10.5	40.1 ± 25.9	33.6 ± 40.2	**89.1 ± 10.0**	76.6 ± 14.1	77.4 ± 16.0	67.2 ± 21.3	**85.9 ± 11.8**
10m	90.5 ± 3.8	33.9 ± 25.2	54.7 ± 44.4	**93.8 ± 2.2**	56.2 ± 20.6	36.8 ± 20.7	32.9 ± 11.3	**61.3 ± 11.3**
12m	**70.8 ± 15.2**	10.9 ± 18.9	17.2 ± 28.0	58.6 ± 11.8	24.0 ± 10.5	4.0 ± 5.3	3.2 ± 3.8	**35.9 ± 8.1**
7m8m	18.8 ± 3.1	0.8 ± 1.4	0.0 ± 0.0	**25.0 ± 15.1**	1.6 ± 1.6	2.4 ± 2.6	0.0 ± 0.0	**28.1 ± 22.0**
8m9m	15.8 ± 3.3	1.6 ± 1.6	0.0 ± 0.0	**19.6 ± 6.0**	3.1 ± 3.8	3.1 ± 3.1	2.3 ± 2.6	**4.7 ± 2.7**
10m11m	**45.3 ± 11.1**	0.8 ± 1.4	0.0 ± 0.0	42.2 ± 7.2	19.7 ± 8.9	2.4 ± 1.4	4.0 ± 3.4	**29.7 ± 15.4**
10m12m	1.0 ± 1.5	0.0 ± 0.0	0.0 ± 0.0	**1.6 ± 1.6**	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**1.6 ± 1.6**
13m15m	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**2.3 ± 2.6**	0.6 ± 1.3	0.0 ± 0.0	0.0 ± 0.0	**1.6 ± 1.6**

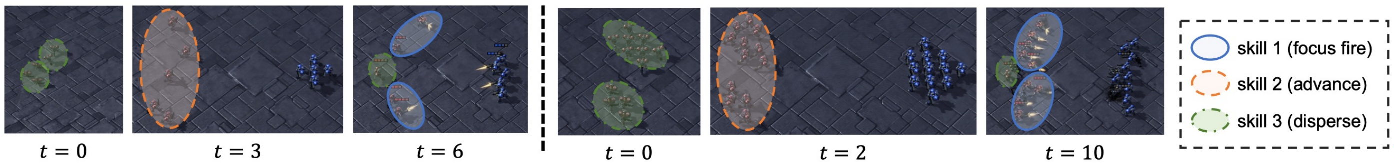
Task	Medium-Expert: BC-best	Medium-Expert: UPDeT-l	Medium-Expert: UPDeT-m	Medium-Expert: ODIS (ours)	Medium-Replay: BC-best	Medium-Replay: UPDeT-l	Medium-Replay: UPDeT-m	Medium-Replay: ODIS (ours)
Source Tasks								
3m	67.7 ± 23.7	50.1 ± 23.9	**85.2 ± 17.9**	73.6 ± 22.0	81.1 ± 8.8	27.3 ± 25.9	41.4 ± 20.1	**83.6 ± 14.0**
5m6m	**31.3 ± 6.3**	2.3 ± 2.6	1.6 ± 1.6	9.4 ± 2.2	**25.0 ± 3.1**	0.8 ± 1.4	0.8 ± 1.4	16.6 ± 4.7
9m10m	26.0 ± 13.9	27.7 ± 24.1	24.3 ± 18.7	**31.3 ± 14.5**	33.4 ± 13.1	2.3 ± 4.1	0.8 ± 1.4	**34.4 ± 8.0**
Unseen Tasks								
4m	81.3 ± 18.9	41.0 ± 8.0	43.9 ± 39.0	**82.8 ± 13.5**	**61.5 ± 9.0**	23.4 ± 15.5	35.9 ± 12.6	55.6 ± 14.5
5m	74.0 ± 2.9	65.7 ± 10.1	33.6 ± 40.2	**82.8 ± 17.7**	75.0 ± 24.2	54.7 ± 23.5	61.7 ± 20.3	**96.1 ± 4.1**
10m	78.1 ± 6.7	39.8 ± 20.1	32.8 ± 38.1	**82.8 ± 16.8**	82.4 ± 8.2	8.6 ± 8.7	11.0 ± 7.8	**84.4 ± 15.1**
12m	64.8 ± 24.3	9.4 ± 7.9	9.4 ± 8.6	**81.3 ± 20.6**	83.4 ± 4.5	2.3 ± 4.1	2.3 ± 2.6	**84.4 ± 6.6**
7m8m	13.3 ± 4.5	4.0 ± 4.2	2.3 ± 4.1	**15.6 ± 4.4**	7.3 ± 6.4	2.3 ± 2.6	1.6 ± 2.7	**9.4 ± 2.2**
8m9m	10.2 ± 4.6	5.6 ± 4.8	9.5 ± 8.6	**10.9 ± 4.7**	11.5 ± 3.9	0.8 ± 1.4	0.8 ± 1.4	**11.7 ± 8.7**
10m11m	26.6 ± 4.7	8.0 ± 12.2	11.8 ± 8.1	**33.6 ± 8.9**	**46.8 ± 6.6**	2.3 ± 4.1	0.8 ± 1.4	35.9 ± 5.2
10m12m	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**1.6 ± 1.6**	1.6 ± 2.7	0.0 ± 0.0	0.0 ± 0.0	**2.3 ± 1.4**
13m15m	0.8 ± 1.4	0.0 ± 0.0	0.0 ± 0.0	**2.3 ± 2.6**	1.6 ± 1.6	0.0 ± 0.0	0.0 ± 0.0	**2.4 ± 1.4**

Semantics of Discovered Coordination Skills

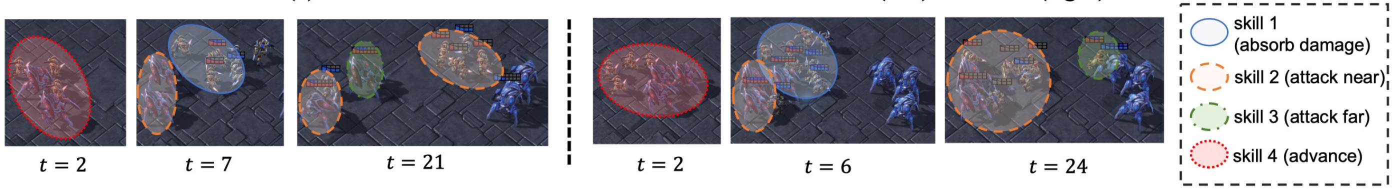


(a) Coordination skill usages in the marine-easy task set

(b) Coordination skill usages in the stalker-zealot task set



(c) Visualization of different coordination skills in 5m (left) and 12m (right)



(d) Visualization of different coordination skills in 2s3z (left) and 3s4z (right)

Effective skills can be summarized at similar timesteps from different tasks!

Discovering Generalizable Multi-agent Coordination Skills from Multi-task Offline Data

Thank you!

Presented by **Fuxiang Zhang**

Published as a conference paper at ICLR 2023

https://github.com/LAMDA-RL/ODIS