Coach-Player Multi-agent Reinforcement Learning for Dynamic Team Composition

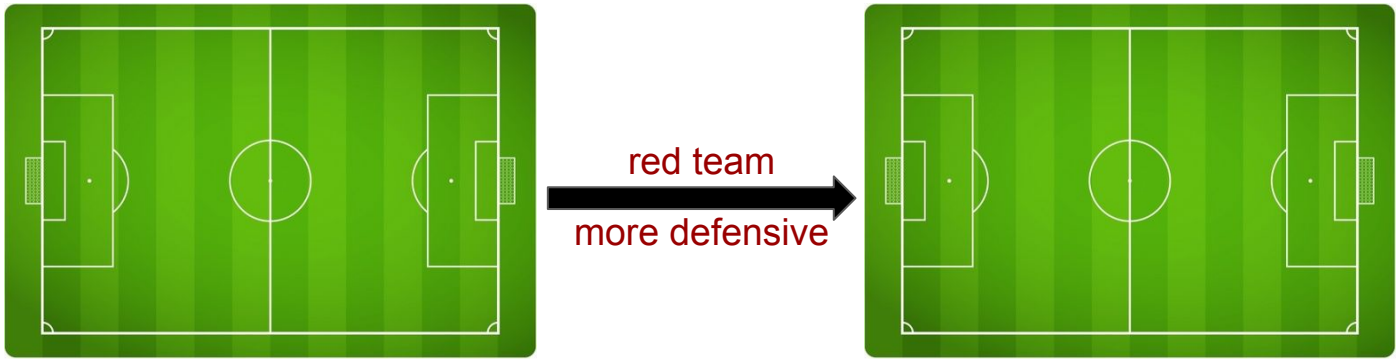
Bo Liu[1], Qiang Liu[1], Peter Stone[1], Animesh Garg[2,3], Yuke Zhu[1,3], Animashree Anandkumar[3,4]

[1]The University of Texas at Austin, [2]University of Toronto,
[3]Nvidia, [4]California Institute of Technology

38th International Conference on Machine Learning, 2021

Motivation

In real world multi-agent systems, agents with different ***capabilities*** may join or leave. So the optimal team strategy can vary according to the ***dynamic team composition***.



Problem: how to coordinate multi-agent systems that have dynamic team composition?

Notation

We model the cooperative multi-agent tasks using the *Decentralized Partially Observable Markov Decision Process* (Dec-POMDP) with entities [1] and extend the representation to include characteristics for each agents (e.g. abilities).

[1] de Witt et al., 2019. Multi-agent common knowledge reinforcement learning.

Notation

A Dec-POMDP with entities and characteristics can be described as a tuple:

$$(S, U, O, P, R, \mathcal{E}, \mathcal{A}, \mathcal{C}, m, \Omega, \rho, \gamma),$$

where

$\mathcal{E}$: the set of all entities (agents/other landmarks)

S : state space ($s = \{s^e | e \in \mathcal{E}\} \in S$)

U : joint action space ($u = \{u^a | a \in \mathcal{A}\} \in U$)

O : observation space ($o^a = \{s^e | m(a, e) = 1\} \in O$)

P : the transition dynamics ($s' \sim P(s, u; c)$)

R : the reward function ($r \sim R(s, u; c)$)

$\mathcal{A}$: the set of all agents ($\mathcal{A} \subseteq \mathcal{E}$)

$\mathcal{C}$: characteristics space of the entity ($c^e \in \mathcal{C}$)

Ω : scenario space ($c = \{c^e | e \in \mathcal{E}\} \in \Omega$)

m : observability function ($m : \mathcal{A} \times \mathcal{E} \to \{0, 1\}$)

ρ : scenario distribution ($c \sim \rho(c)$)

γ : the discount factor

Dec-POMDP Objective

The learning of objective of a Dec-POMDP is to maximize the cumulative return:

$$G = \mathbb{E}_{s_0,u_0,s_1,u_1,\ldots}[\sum_{t=0}^{\infty} \gamma^t r_t].$$

Under the partial observability assumption, the history of observation-action pairs is often encoded with a compact vector representation in place of the state. Let $Q_\theta^{\text{tot}}(\tau_t, u_t; c)$ represent the team's action-value function given history $\tau_t = \{\tau_t^a = (o_0^a, u_0^a, \ldots o_t^a) | \text{agent } a\}$. Then normally we minimize the following Bellman error

$$\mathcal{L}(\theta) = \mathbb{E}_{(c,\tau_t,u_t,r_t,\tau_{t+1})\sim\mathcal{D}}\left[\left(r_t + \gamma \max_{u'} Q_{\bar{\theta}}^{\text{tot}}(\tau_{t+1}, u'; c) - Q_\theta^{\text{tot}}(\tau_t, u_t; c)\right)^2\right].$$

Here, $\mathcal{D}$ is a replay buffer that stores previously generated off-policy data. $Q_{\bar{\theta}}^{\text{tot}}$ is the target network parameterized by a delayed θ copy of for stability.

Related work

- **Centralized training with decentralized execution (CTDE).** CTDE assumes agents execute independently but uses the global information for training [2,3,4,5]. Unlike prior methods, we study how to incorporate global information for teams with dynamic compositions.
- Prior works in transfer and curriculum learning transfer policies of small teams for larger teams [6,7,8,9,10]. While these works mainly focus on homogeneous agents, we focus on heterogeneous agents (e.g. agents with different characteristics) and dynamic teams.
- **Ad hoc teamwork.** Standard ad hoc teamwork research focuses on the single ad hoc agent and assumes no control over the teammates [11, 12]. Recent works study how to coordinate based on pre-specified communication protocols [13,14]. We study how to coordinate the entire team with learnable strategies.

Global Information

As the team composition is subject to change, it is important for the team to be aware of any scenario (e.g. composition) changes as soon as they happen.

Example: consider a single state (bandit) problem where the reward is defined as

$$R(u;c) = \max_a c^a u^a + 1 - \sum_a u^a, \quad u^a \in \{0,1\}$$

In other words, the team is punished if more than 1 agent performs the task and if only 1 agent performs the task, the reward is proportional to its characteristics c^a. Clearly, if the team composition (e.g. the scenario c) is changing, the optimal strategy requires knowing the global information.

Coach-Player Multi-agent RL (COPA)

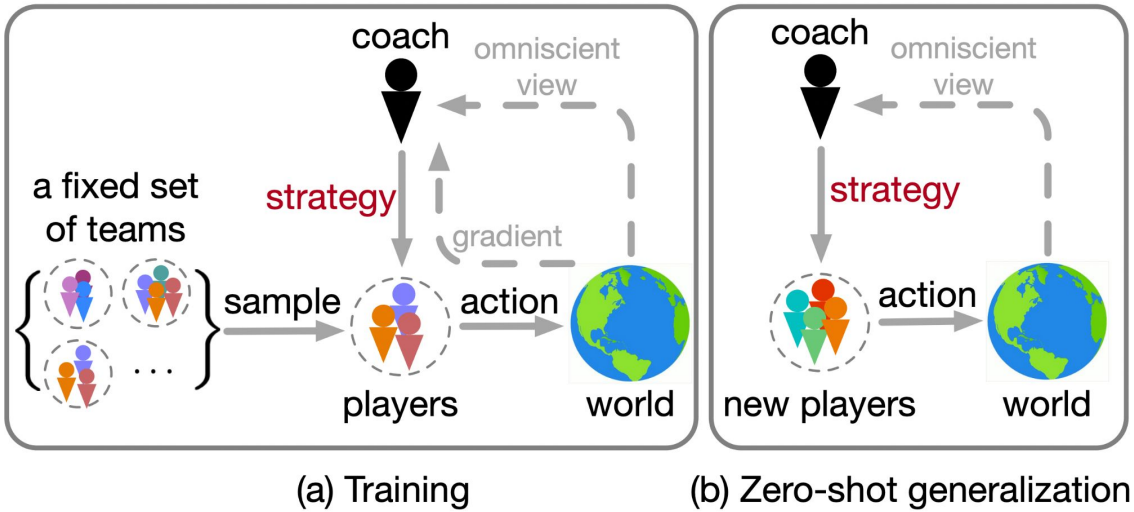
Knowing that global information can be necessary for dynamic team composition, we propose the COach-PlAyer multi-agent reinforcement learning method (COPA).

Specifically, COPA consists of

- a coach that has *omniscient* view of the world, and distributes strategies to players occasionally.
- players (a.k.a agents) who have *partial* views of the world and coordinate with each other based on the received strategies from the coach.

Coach-Player Multi-agent RL (COPA)

Therefore, in training, we sample different scenarios and train the coach and players to cooperatively perform the task. During testing, the team might encounter unseen team composition and the hope is that COPA can generalize in a zero-shot fashion.



(a) Training (b) Zero-shot generalization

Coach-Player Multi-agent RL (COPA)

Coach:

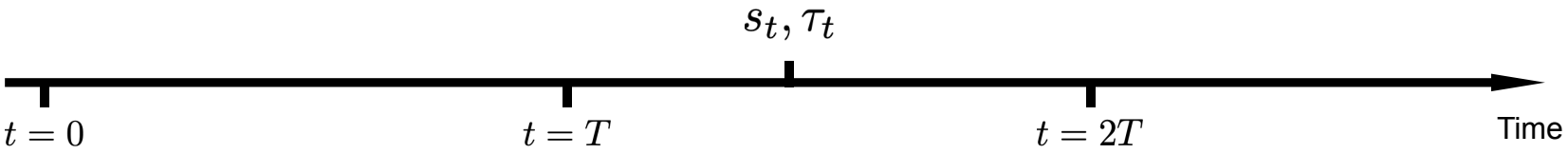
- Compute strategies every T steps (for example, at $t = T$):

$$z_T^a \sim \mathcal{N}(\mu^a, \Sigma^a), \quad (\mu = \{\mu^a \mid a \in \mathcal{A}\}, \Sigma = \{\Sigma^a \mid a \in \mathcal{A}\}) = f_\phi(s_T; c)$$

Players:

- Condition on the most recent strategy and the local history, performs an action:

$$u_t^a \sim Q_\theta^a(\tau_t^a, z_{\hat{t}}; c^a), \quad \hat{t} = \max_{t'} \{t' \equiv 0 \pmod{T},\ t' \leq t\}$$



Coach-Player Multi-agent RL (COPA)

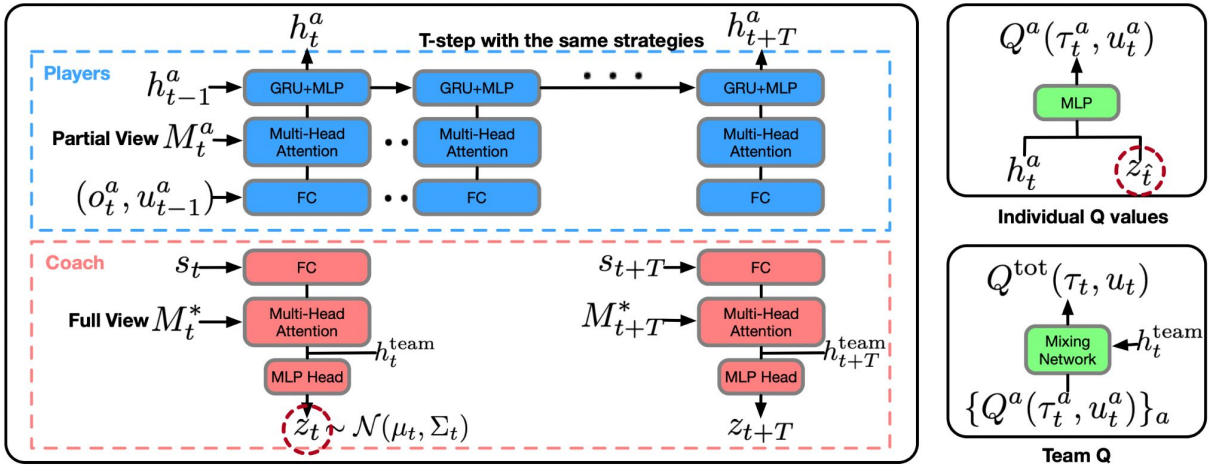
The RL objective therefore becomes:

$$\mathcal{L}_{\mathrm{RL}}(\theta, \phi) = \mathbb{E}_{(c,\tau_t,u_t,r_t,s_{\hat{t}},s_{t\hat{+}1})\sim\mathcal{D}}\left[\left(r_t + \gamma \max_{u'} Q^{\mathrm{tot}}_{\bar{\theta}}(\tau_{t+1}, u' | z_{t\hat{+}1}; c) - Q^{\mathrm{tot}}_{\theta}(\tau_t, u_t \mid z_{\hat{t}}; c)\right)^2\right],$$

where $z_{\hat{t}} \sim f_{\phi}(s_{\hat{t}}; c)$, $z_{t\hat{+}1} \sim f_{\bar{\phi}}(s_{t\hat{+}1}; c)$, and $\bar{\phi}$ denotes the parameters of the target network for the coach's strategy predictor f.

COPA (model)

To deal with an unknown number of heterogeneous agents, we adopt the multi-head attention model from a previous multi-agent RL work REFIL [15]. On top of that, we add a coach module that receives the full view of the world and broadcasts strategies. The team action-value function Q^{tot} is mixed by a mixing network using individual Q^a.



COPA (regularization)

Inspired by recent works [16,17] that apply variational inference to regularize the learning of a latent space in reinforcement learning, we propose a variational objective to ensure that the received strategy of an agent is identifiable from its future behavior. Denote $\zeta_t^a = (o_{t+1}^a, u_{t+1}^a, o_{t+2}^a, u_{t+2}^a, \ldots, o_{t+T-1}^a, u_{t+T-1}^a)$, then we maximize the mutual information

$$I(z_t^a; \zeta_t^a, s_t) \geq \mathbb{E}_{s_t, z_t^a, \zeta_t^a} \left[\log q_\xi(z_t^a | \zeta_t^a, s_t) \right] + H(z_t^a | s_t).$$

We further adopt the Gaussian factorization for the variational variable q_ξ as in [3]:

$$q_\xi(z_t^a | \zeta_t^a, s_t) \propto q_\xi^{(t)}(z_t^a | s_t, u_t^a) \prod_{k=t+1}^{t+T-1} q_\xi^{(k)}(z_t^a | o_k^a, u_k^a).$$

COPA (communication frequency)

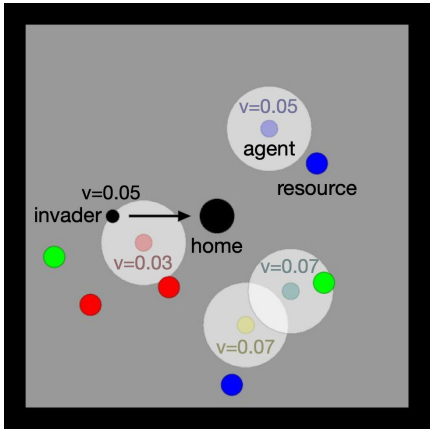
So far, we assumed the coach broadcasts the strategies periodically. In practice, this may incur communication costs. So we design a simple method to adaptively control the communication frequency: the coach decides whether or not to broadcast a new strategy to an agent based on the ℓ^2 distance between the new and old strategies.

$$\tilde{z}_t^a = \begin{cases} z_t^a \sim f_\phi(s, c) & \text{if } ||z_t^a - z_{\text{old}}^a||_2 \geq \beta \\ z_{\text{old}}^a & \text{otherwise.} \end{cases}$$

Remark: Note that the threshold β is chosen *after* the model is trained. By adjusting β, one can easily achieve different communication frequencies.

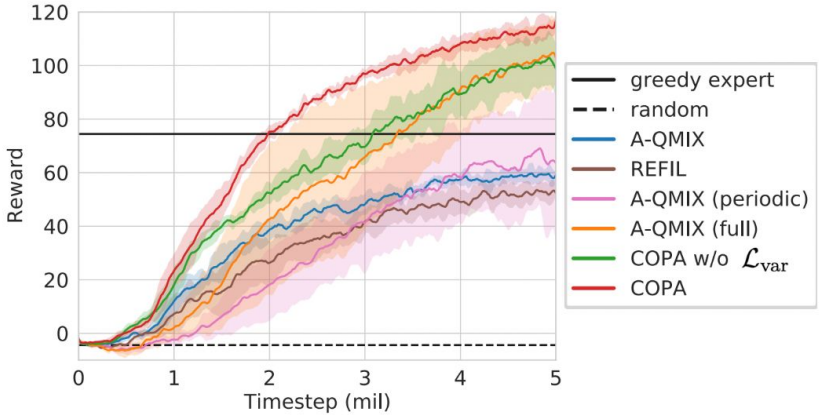
Experiment (resource collection)

In resource collection, a team of 2-4 agents needs to collect as much resource as possible and bring them home. There are three types of resources (r, g and b) spread out the world that can respawn once being collected. For each agent, its characteristics is $(c_r^a, c_g^a, c_b^a, v^a)$, where the first three indicate how efficiently an agent collects the three resources and $v^a$is the speed of the agent.

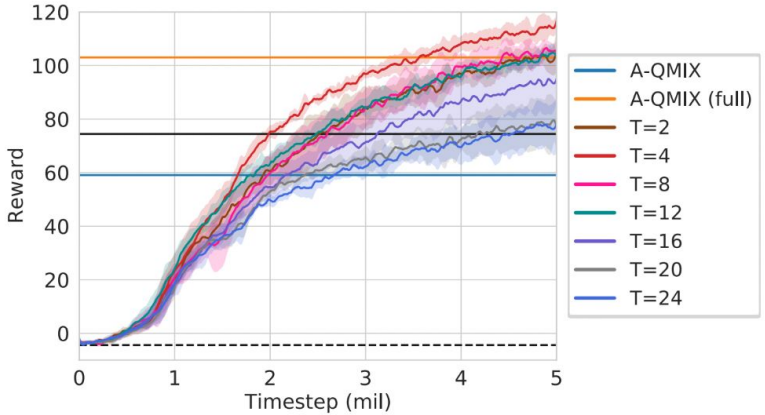


Experiment (resource collection)

We compare against baseline methods that do not have any global information (e.g. A-QMIX and REFIL [15]), which are essentially COPA without the coach. We also compare against their variants that periodically receive the full view of the environment (A-QMIX (periodic)) or keep receiving the full view (A-QMIX (full)).



(a) Comparison with baselines



(b) Ablations on T

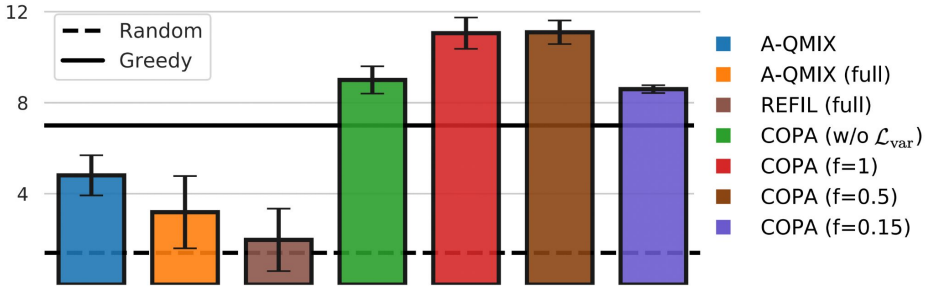
Experiment (resource collection)

We also evaluates COPA's zero-shot generalization to environments having 5, 6 or a varying number of agents, with different levels of communication frequency.

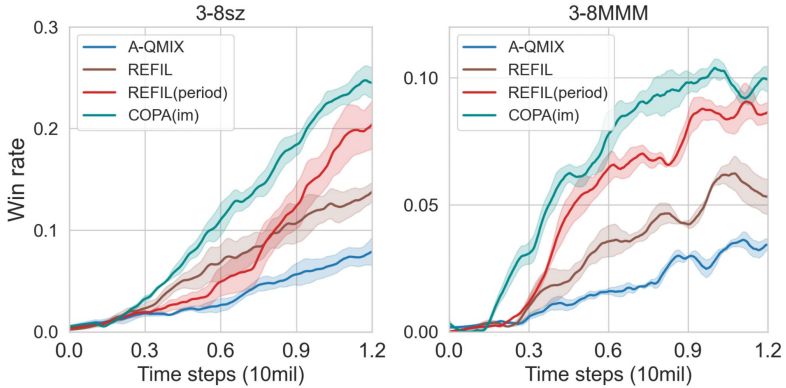
Method	Env. ($n = 5$)	Env. ($n = 6$)	Env. (varying n)	f
Random Policy	6.9	10.4	2.3	N/A
Greedy Expert	115.3	142.4	71.6	N/A
REFIL	90.5±1.5	109.3±1.6	61.5±0.9	0
A-QMIX	96.9±2.1	115.1±2.1	66.2±1.6	0
A-QMIX (periodic)	93.1±20.4	104.2±22.6	68.9±12.6	0.25
A-QMIX (full)	157.4±8.5	179.6±9.8	114.3±6.2	1
COPA ($\beta = 0$)	175.6±1.9	203.2±2.5	124.9±0.9	0.25
COPA ($\beta = 2$)	174.4±1.7	200.3±1.6	122.8±1.5	0.18
COPA ($\beta = 3$)	168.8±1.7	195.4±1.8	120.0±1.6	0.13
COPA ($\beta = 5$)	149.3±1.4	174.7±1.7	104.7±1.6	0.08
COPA ($\beta = 8$)	109.4±3.6	130.6±4.0	80.6±2.0	0.04

Experiment (rescue game & SMAC)

We also conduct experiments on a rescue game and the starcraft micromanagement challenge (SMAC). COPA consistently outperforms baseline methods.



Rescue Game



SMAC

Reference

[1] de Witt et al., 2019. Multi-agent common knowledge reinforcement learning.
[2] Sunehag et al., 2017. Value-decomposition networks for cooperative multi-agent learning.
[3] Rashid et al., 2018. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning.
[4] Mahajan et al., 2019. Multi-agent variational exploration.
[5] Son et al., 2019. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning.
[6] Carion et al., 2019. A structured prediction approach for generalization in cooperative multi-agent reinforcement learning.
[7] Shu & Tian, 2019. Mind-aware multi-agent management reinforcement learning.
[8] Agarwal et al., 2019. Learning transferable cooperative behavior in multi-agent teams.
[9] Wang et al., 2020. From few to more: Large-scale dynamic multiagent curriculum learning.
[10] Long et al., 2020. Evolutionary population curriculum for scaling multi-agent reinforcement learning.
[11] Genter et al., 2011. Collaboration in ad hoc teamwork: ambiguous tasks, roles, and communication.
[12] Barrett & Stone, 2012. Communicating with unknown teammates.
[13] Grizou et al., 2016. Collaboration in ad hoc teamwork: ambiguous tasks, roles, and communication.
[14] Mirsky et al., 2020. A penny for your thoughts: The value of communication in ad hoc teamwork.
[15] Iqbal et al., 2020. Attention and imagination for dynamic multi-agent reinforcement learning.
[16] Rakelly et al., 2019. Efficient off-policy meta-reinforcement learning via probabilistic context variables.
[17] Wang et al., 2020. Multiagent reinforcement learning with emergent roles.

Thanks!


Bo Liu


Qiang Liu


Peter Stone


Animesh Garg


Yuke Zhu


Animashree Anandkumar

Paper: