TOWARDS OPEN AD HOC TEAMWORK USING GRAPH-BASED POLICY LEARNING

Arrasy Rahman, Niklas Höpner, Filippos Christianos, Stefano Albrecht

In 38th International Conference on Machine Learning 2021

THE UNIVERSITY of EDINBURGH informatics

Autonomous Agents Research Group

OPEN AD HOC TEAMWORK

Ad hoc teamwork :

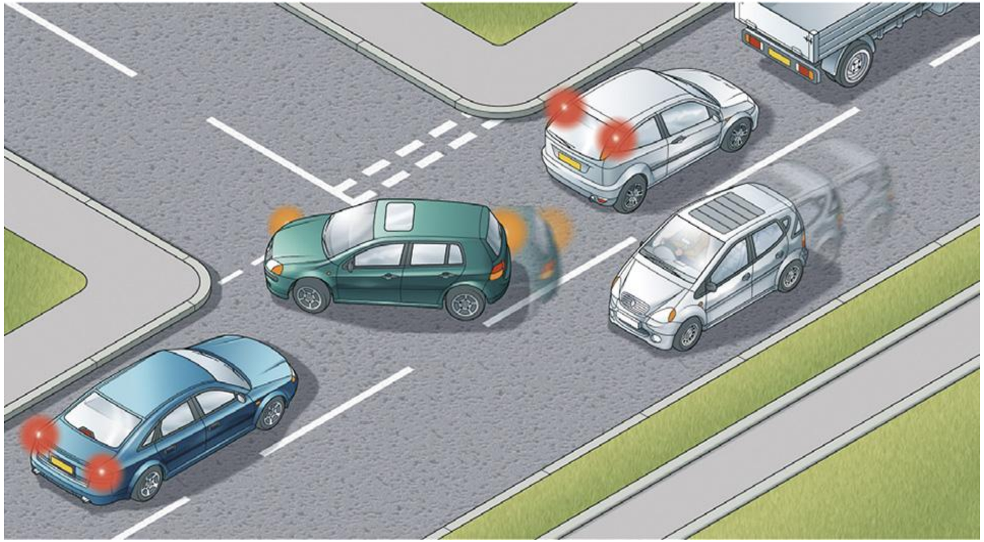
- Control a single agent (**learner**)
- Maximize returns in the presence of other agents without **prior coordination mechanisms**

Open multiagent systems :

- Agents may leave or enter the environment anytime.
- Number of agents may change between timesteps.

CHALLENGES FOR OPEN AD HOC TEAMWORK

1. Adaptation to different teammate policies
2. Adaptation to changing team sizes
3. Handling variable observation sizes



LEARNING OBJECTIVE

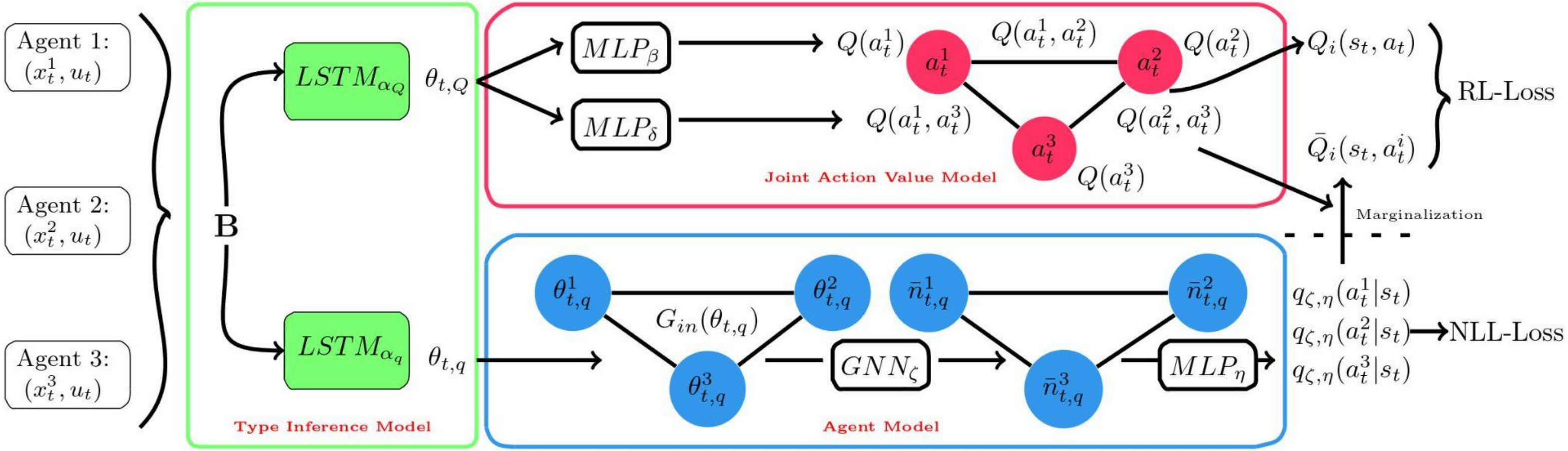
Find an optimal policy for the learner, $\boldsymbol{\pi}^{i,*}$, characterized by:

$$\forall \pi^i, s, a^i, \bar{Q}_{\pi^{i,*}}(s, a^i) \geq \bar{Q}_{\pi^i}(s, a^i)$$

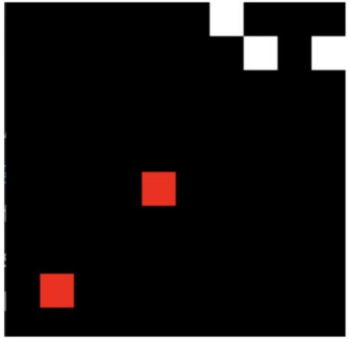
with,

$$\bar{Q}_{\pi^i}(s, a^i) = \mathbb{E}_{a^i_t \sim \pi^i, a^{-i}_t \sim \boldsymbol{\pi}^{-i}_t, P}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \middle| s_0 = s, a^i_0 = a^i\right]$$

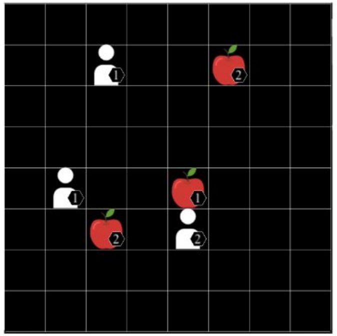
GRAPH-BASED POLICY LEARNING (GPL)



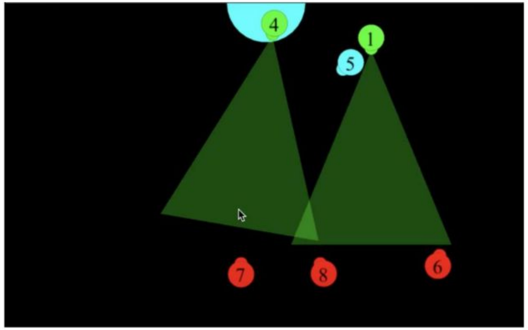
EXPERIMENTS: ENVIRONMENTS



(a) Wolfpack



(b) Level-based foraging



(c) FortAttack

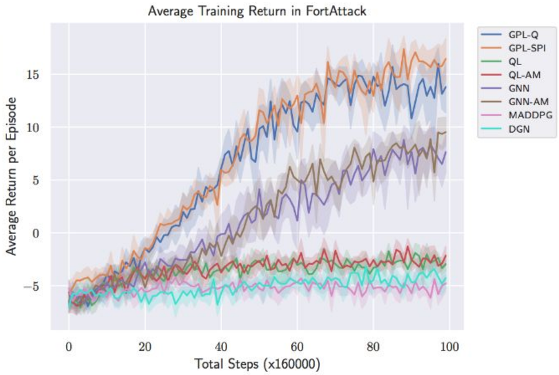
EXPERIMENTS: RESULTS

Results under training setup

- Open process changes number of teammates between timesteps
- Up to 2 teammates at any timestep

(a) Results in Level-based foraging

(b) Results in Wolfpack



(c) Results in FortAttack

EXPERIMENTS: GENERALIZATION RESULTS

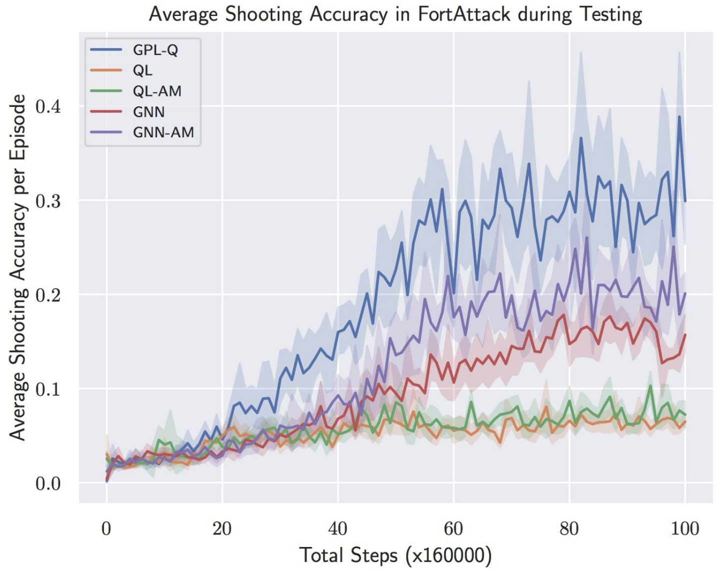
Results under generalization setup

- Open process restricts teammates to up to 4 agents at any timestep
- Only evaluate policies that delivered best performance during training

Env.	GPL-Q	GPL-SPI	QL	QL-AM	GNN	GNN-AM	DGN	MADDPG
LBF	**2.32±0.22**	**2.40±0.16***	1.41±0.14	1.22±0.29	2.07±0.13	1.80±0.11	0.64 ± 0.9	0.91 ± 0.10
Wolf.	**36.36±1.71***	**37.61±1.69***	20.57±1.95	14.24±2.65	8.88±1.57	30.87±0.95	2.18 ± 0.66	19.20 ± 2.22
Fort.	**14.20±2.42***	**16.82±1.92***	-3.51±0.60	-3.51±1.51	7.01±1.63	8.12±0.74	-5.98 ± 0.82	-4.83 ± 1.24

WHY GPL WORKS : AN ANALYSIS IN FORTATTACK

- Which GPL component is responsible for its performance?
- How does this component yield high returns?



ANALYSIS IN FORTATTACK : SHOOTING METRIC

- Evaluate several shooting-related metrics and measure correlation with yielded returns
- Among all metrics,

$$\bar{Q}_{j,k} = \frac{\sum_{a^k} Q_{\delta}^{j,k}(a^j = \text{shoot}, a^k | s)}{|A^k|}$$

, by far has the highest correlation with returns.

ANALYSIS IN FORTATTACK : ANALYSIS OF METRIC

- Strong correlation when *j* is a defender and *k* is an attacker
- MLP_δ learns that :

 “If *k* is an attacker inside j’s (any defender) shooting range $\rightarrow$ High shooting values for *j* shooting *k.*”

- MLP_δ enables reuse of knowledge

ANALYSIS IN FORTATTACK : VALUE LEARNING IN OTHER BASELINES

- Can other baselines learn the effects of other agents' actions towards the learner? Turns out they can't
- Learning process in baselines
  - Learner must successfully shoot attackers itself to increase the value of shooting
  - Shooting well-trained opposition is difficult
  - Baselines do not learn the value of other teammates' actions
  - See our additional experiments in appendix.

Towards Open Ad Hoc Teamwork Using Graph-based Policy Learning

https://arxiv.org/abs/2006.10412

Contributions:

1. We present the first approach to solve open ad hoc teamwork.
2. We demonstrate the importance of GNNs for handling environment openness.
3. We empirically proved that modelling the effects of other teammates' actions yields higher returns in open ad hoc teamwork.

In 38th International Conference on Machine Learning 2021

THE UNIVERSITY of EDINBURGH informatics
Autonomous Agents Research Group