Multi-agent deep reinforcement learning in mobile robotics

Comité de suivi de 1ère année

Maxime Toquebiau

ECE Paris, Research Center, 37 quai de Grenelle, 75015, Paris, France

Sorbonne Université, Institut des Systèmes Intelligents et de Robotique, ISIR, 4 place Jussieu, 75005 Paris, France

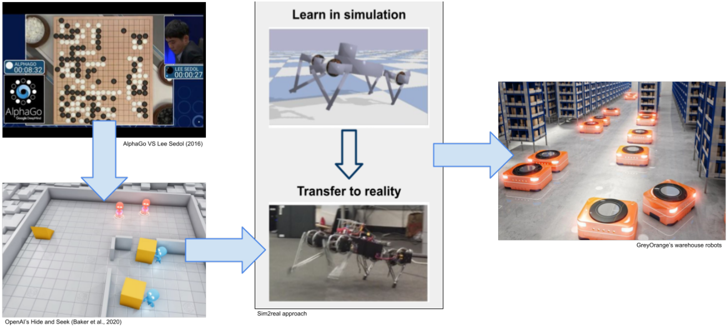
February 3rd, 2022

Multi-agent deep reinforcement learning in mobile robotics

Multi-agent deep reinforcement learning in mobile robotics

OR

Building a path to apply multi-agent deep reinforcement learning to mobile robotics



Our goal:

Perform a task in the real world with a multi-robot system.

⇒ The environment is partially observable.
⇒ We need to:

- be able to interpret our agents' strategy,
- interact with them.

Our proposed path:

1) Emergent communication

⇩

2) Learning an existing language

⇩

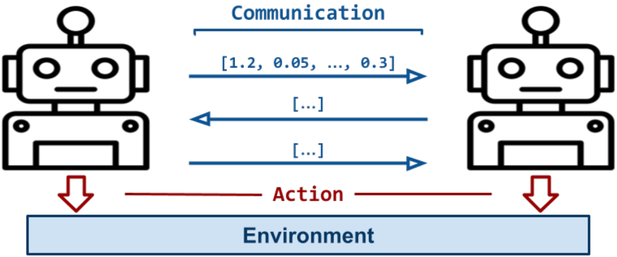
3) Multi-agent credit assignment

⇩

4) Macro-actions as a sim2real approach

Context:

- Partial observability prevents agents from having crucial information about the environment.



Proposed approach:

- Agents can overcome this issue by communicating their local observations.

Context:

- Need to *interact with our agents* and *understand their actions*.
- Emergent languages are very difficult to interpret (Lazaridou and Baroni, 2020).
- A discrete language is a way to understand the world and construct logical thoughts (Vygotsky, 1934).

Proposed approach:

- Learn a pre-defined language made of discrete tokens.

Type	Token
Entities	AGENT PACKAGE
Locations	NORTH SOUTH EAST WEST DELIVERY_AREA
Verbs	GOING_TO NEED_HELP PUSH

Possible tokens used in a language designed for a delivery task.

Context:

- The global reward does not necessarily reflect the quality of each agent's actions.
- The effect of communication must be accounted for in the quality of an agent's policy.

Proposed approach:

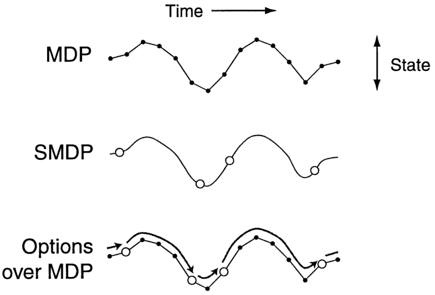
- Explore credit assignment methods in MADRL.
- Find ways to include communication in this process.

Context:

- DRL is difficult to perform in the real world.
- *Reality gap*: set of all differences between simulation and reality.

Proposed approach:

- Use macro-actions as a simulation-to-reality (sim2real) tool.
- On the macro level, the reality gap is thinner.



Macro-actions (or options) as defined by Sutton et al. (1999).

What has been done this year:

1) Literature review
2) Creating a simulated environment
3) Training methods from the literature

1) Literature review

Value-based methods:

$$V_\pi(s_t) = E_\pi[G_t | s_t]$$

$$Q_\pi(a_t | s_t) = E_\pi[G_t | a_t, s_t]$$

- **Deep Q-Network (DQN)** (Mnih et al., 2013, 2015)
- Prioritized Experience Replay (Schaul et al., 2016)
- Double DQN (van Hasselt et al., 2016)
- Dueling DQN (Wang et al., 2016)
- Noisy DQN (Fortunato et al., 2017)
- Recurrent DQN (Kapturowski et al., 2018)

Policy-based methods:

$$\pi(s) = a$$

- **Deep Deterministic Policy Gradient (DDPG)** (Lillicrap et al., 2015)
- Trust Region Policy Gradient (TRPO) (Schulman et al., 2015)
- **Proximal Policy Optimisation (PPO)** (Schulman et al., 2017)
- Twin Delayed DDPG (TD3) (Fujimoto et al., 2018)
- **Soft Acotr-Critic (SAC)** (Haarnoja et al., 2018)

Model-based methods:

Predicting the next states of the environment.
Use that for planning.

- **AlphaGo** (Silver et al., 2016)
- AlphaZero (Silver et al., 2017)
- World Models, (Ha and Schmidhuber, 2018)
- MuZero (Schrittwieser et al., 2019)
- PlaNet (Hafner et al., 2018)
- Dreamer (Hafner et al., 2019)

Centralised training, decentralised execution (CTDE):

- **Multi-agent DDPG (MADDPG)** (Lowe et al., 2017)

Independent learning:

- Independent DQN (Tampuu et al., 2017)
- Independent PPO (Schroeder de Witt et al., 2020)

Centralised execution:

- Deep Coordination Graphs (Böhmer et al., 2020)

Multi-agent Credit Assignment:

Dividing the reward given to the MAS, to match each agent's actions.

- Wonderful Life Q-function (WLQ) and Aristocrat Utility (Nguyen et al., 2018)
- **COMA** (Foerster et al., 2018)

Value factorisation:

$$\underset{\mathbf{u}}{\operatorname{argmax}}\, Q_{tot}(\boldsymbol{\tau}, \mathbf{u}) = \begin{pmatrix} \operatorname{argmax}_{u^1} Q_1(\tau^1, u^1) \\ \dots \\ \operatorname{argmax}_{u^n} Q_n(\tau^n, u^n) \end{pmatrix}$$

$$Q_{tot}(\boldsymbol{\tau}, \mathbf{u}) = f \begin{pmatrix} Q_1(\tau^1, u^1) \\ \dots \\ Q_n(\tau^n, u^n) \end{pmatrix}$$

- **Value-Decomposition Networks (VDN)** (Sunehag et al., 2018)
- **QMIX** (Rashid et al., 2018)
- Qtran (Son et al., 2019)
- Multiagent Variational Exploration (MAVEN) (Mahajan et al., 2019)
- Weighted QMIX (Rashid et al., 2020)
- QPLEX (Wang et al., 2021)

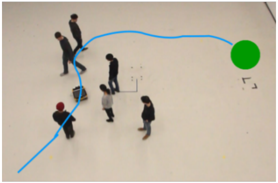
Literature review

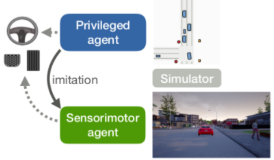
Learning in mobile robotics (1/3)

Navigation using DRL:

- Navigation from pixel input (Zhu et al., 2017; Kahn et al., 2018).
- Planning with other agents (humans and robots) in the environment (Everett et al., 2018, 2021).
- Application on aerial vehicles (Zhang et al., 2016),
- and autonomous vehicles (Shalev-Schwartz et al., 2016; Chen et al., 2020).

Zhu et al., 2017


Everett et al., 2021



Chen et al., 2020

Sim2real:

- Object manipulation through domain randomisation (Tobin et al., 2017; Peng et al., 2018; Andrychowicz et al., 2020).
- Navigation in mobile robotics (Tai et al., 2017; Kang et al., 2019).

Tobin et al., 2017

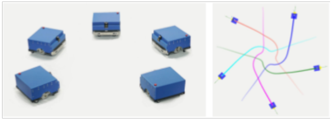

Andrychowicz et al., 2020


Kang et al., 2019

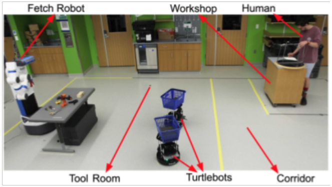
Multi-robot systems:

- Navigation with collision avoidance (Chen et al., 2017; Long et al., 2018; Semnani et al., 2020).
- Tool delivery (Xiao et al., 2020).
- Information gathering (Queralta et al., 2020).

Chen et al., 2017



Long et al., 2018



Xiao et al., 2020

Centralised communication network:

- **CommNet** (Sukhbaatar et al., 2016)
- BiCNet (Peng et al., 2017)
- IC3Net (Singh et al., 2019)
- **Targeted Multi-agent Communication (TarMAC)** (Das et al., 2019)

Decentralised execution:

- **DIAL** (Foerster et al., 2016)
- Emergent grounded compositional language (Mordatch and Abbeel, 2018)
- Attentional Communication (Jiang and Lu, 2018)

Communication in value factorisation:

- Variance based control (Zhang et al., 2019)
- Communication minimisation (Wang et al., 2020)
- Temporal message control (Zhang et al. ,2020)

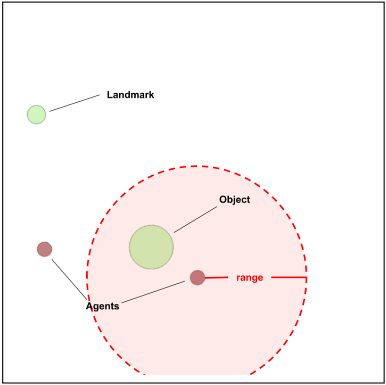
Issues with emergent communication:

- Adds a layer of complexity to the learning process.
- Difficult to measure the efficiency of the learnt language (Lowe et al., 2019).
- Difficult, if not impossible, to interpret (Kottur et al., 2017; Lazaridou and Baroni, 2020)

2) Creating a simulated environment

Multiagent Particle Environment:

- 2D
- Physics-based
- Used in the literature (Lowe et al., 2017; Mordatch and Abbeel, 2018; Jiang and Lu, 2018; Wang et al., 2020)



Coop Push Scenario:

- Agents (weight=0,4kg)
- Movable object (weight=10kg)
- Unmovable landmark
- **Goal:** Move the object on the landmark
- 100 steps maximum per episode
- Partially observable

Observations:

- Own position and velocity
- Position and velocity of agents and objects in observation range
- Position of landmarks in observation range

Actions:

- Discrete: Up/Down/Left/Right/Do nothing
- or Continuous: $[dx, dy], (dx, dy) \in [-1, 1]^2$

Reward:

$$R^t_{dist\_reward} = -dist(object, landmark)^2 - \frac{1}{n_{agent}} \sum_{k=1}^{n_{agent}} dist(agent_k, object) + \rho,$$

with ρ the collision penalty, fixed at -10.

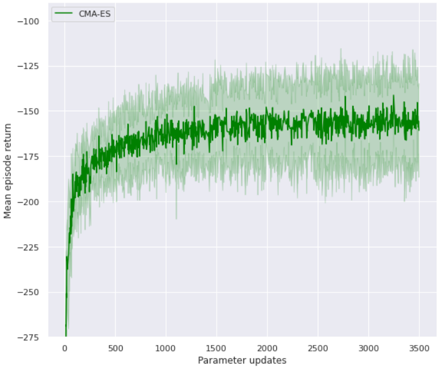
3) Training baselines from the literature

Training baselines from the literature

Chosen models

Model	Paper	Type	Action space
MADDPG	Lowe et al., 2017	Policy-based	Discrete or Continuous
QMIX	Lowe et al., 2017	Value factorisation	Discrete
CMA-ES	Hansen and Ostermeier, 2001	Evolutionary strategy	Discrete or Continuous

Scenario: 2 agents, dist_reward, fully observable



MADDPG: num_episodes=300K, num_updates=150K, lr=0,005, actions=continuous

QMIX: num_episodes=300K, num_updates=150K, lr=0.0005, actions=discrete

CMA-ES: pop_size=18, num_evals=3500, number of episodes per eval = 8, actions=continuous

Future steps

Period	Task
January-February	Find a working reward
February-March	Develop and train emergent communication
March-April	Review Language-Augmented RL
April-June	Develop a system for learning an existing language
June-August	Train agents with existing language
July-October	Publish
September-December	Explore further with credit assignment and macro-actions

One year from now

- Working reward
- A language to teach to artificial agents
- A system for teaching this language
- Proofs that communication improves the agents' performance
- One publication in a conference (NIPS, IROS, CoRL...)

Research

- Meeting every 2 weeks with Jae Yun Jun
- Meeting every 2 months with the whole team

Teaching (100 hours/year)

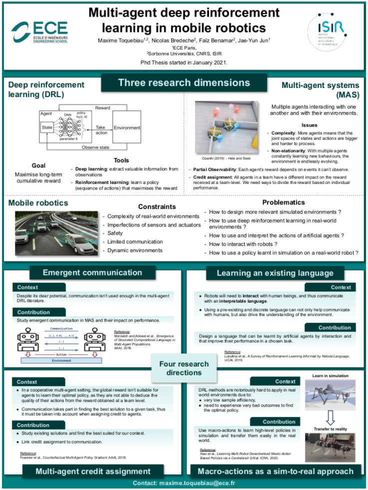
- Programming in C and Python, 1st year ECE students

Presenting my work

- Presentations to students and other researchers
- Poster for the careers fair

Mentoring students in the research minor

- Research project on meta reinforcement learning



Training plan

Technical formation				
Name of the course	**Number of hours**	**Place**	**Course followed**	**Dates**
Robotique Mobile	36	Sorbonne Universités	Yes	12/10/2021 - 03/01/2022
Non-technical formation				
Name of the course	**Number of hours**	**Place**	**Course followed**	**Dates**
Optimiser mes 3 ans de thèse	3	Online		
Scientific writing	12	Online	Yes	20 and 27/05/2021
Anglais - Faire une communication orale	12			

Thank you for your attention.