Multi-agent deep reinforcement learning in mobile robotics

Comité de suivi de 2ème année

Maxime Toquebiau

ECE Paris
Sorbonne Université

January 19th, 2023

Table of Content

Introduction

First year

- Thesis subject
- Literature review
- Simulated environment
- Experimenting with baselines from the literature

CS1 recommendation

- Concentrate on one research direction

Second year

- Research direction: Language Augmented MARL
- Joint Intrinsic Motivation

I. Language-Augmented MARL

Goal and requirements

Goal

Using a pre-defined language to help agents understand their environment and share information efficiently.

Requirements

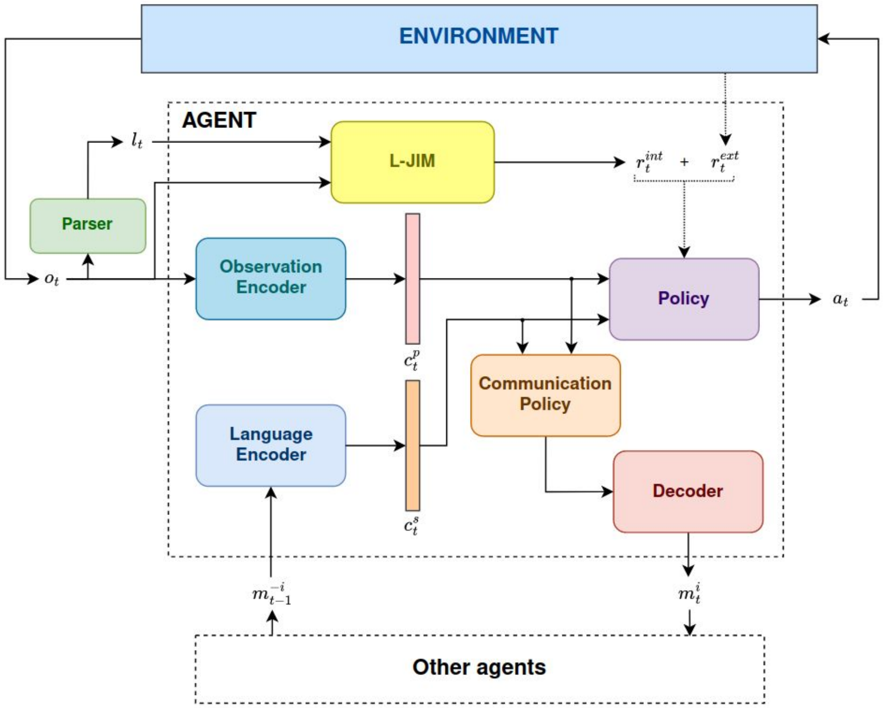
I) A language to teach to agents

II) A way to understand the language

III) A way to generate messages

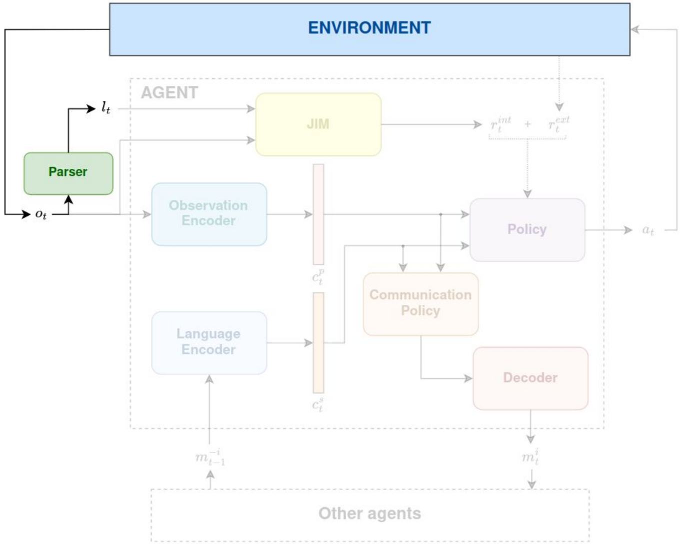
I. Language-Augmented MARL

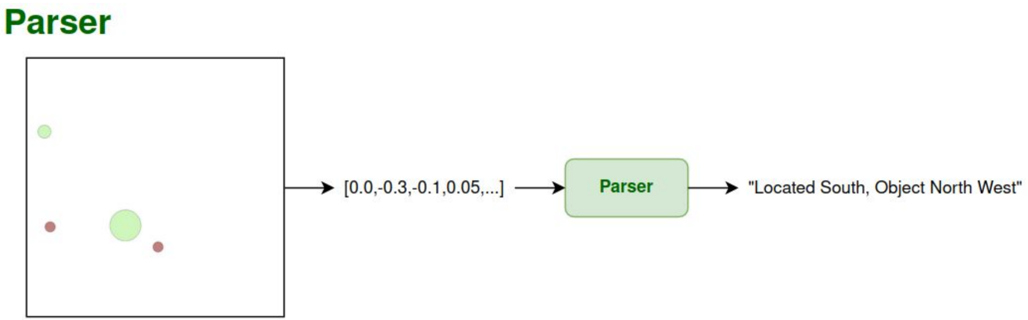
Goal and requirements



I. Language-Augmented MARL

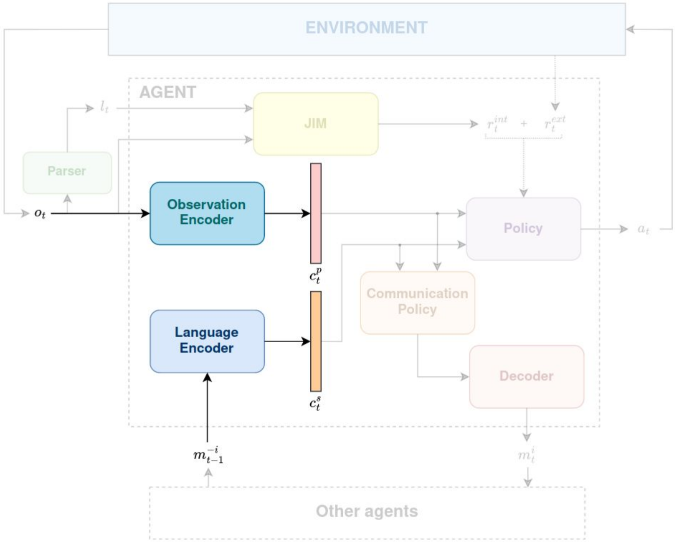
Requirement I: A language to teach to agents

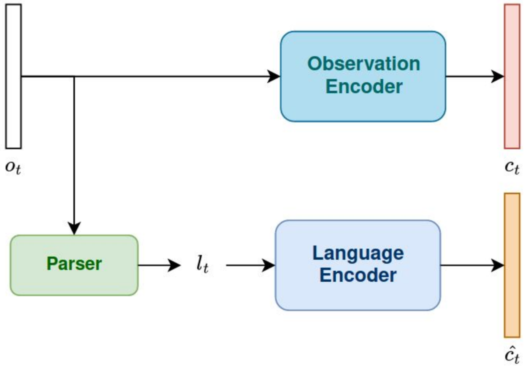




I. Language-Augmented MARL

Requirement II: A way to understand the language





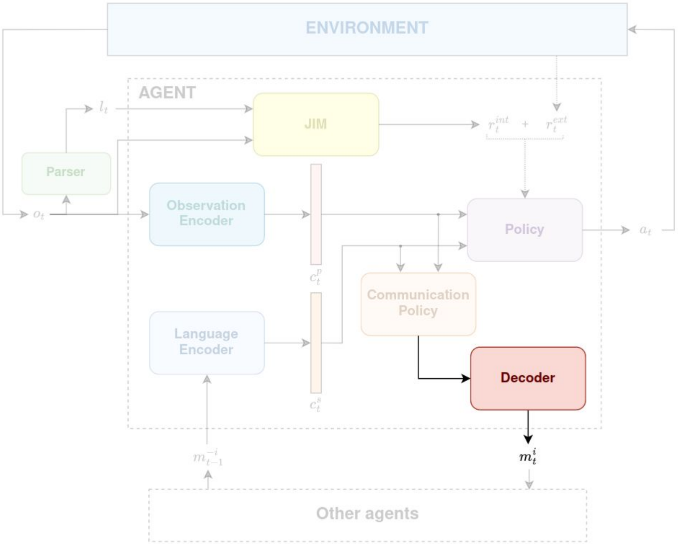
Task: Generate similar encodings of the observation and the description.

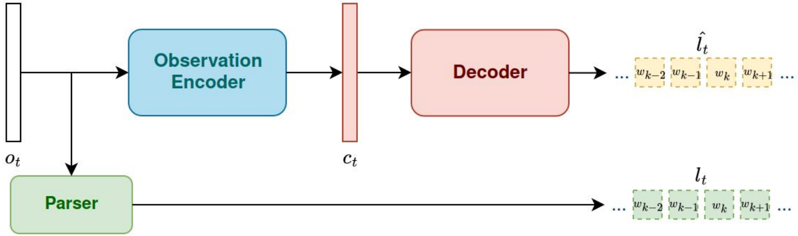
Trained using CLIP[1].

[1] Radford et al., *Learning Transferable Visual Models From Natural Language Supervision*, ICML, 2021

I. Language-Augmented MARL

Requirement III: A way to generate messages

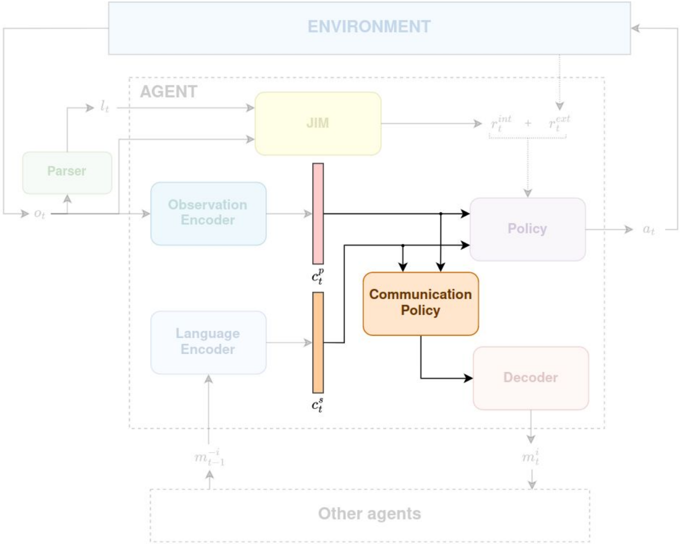


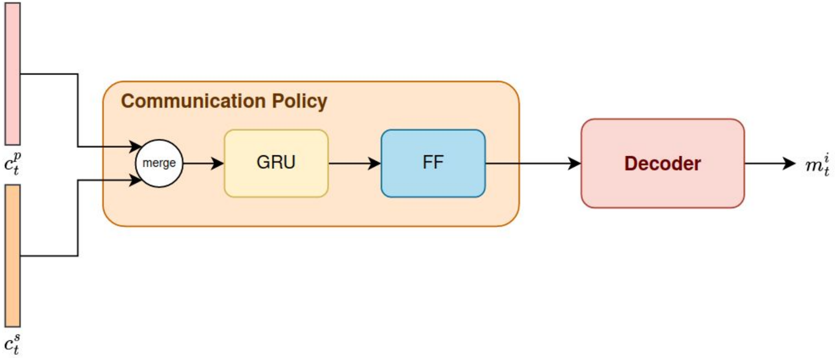


<u>Task:</u> Generating the caption from the observation.

I. Language-Augmented MARL

Requirement III: A way to generate messages





Options for merging:

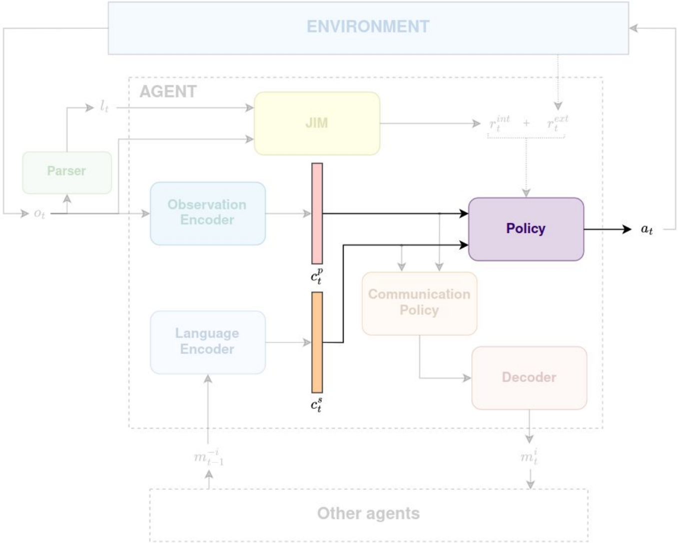
- Concatenation
- Average
- Addition
- Feed forward neural network

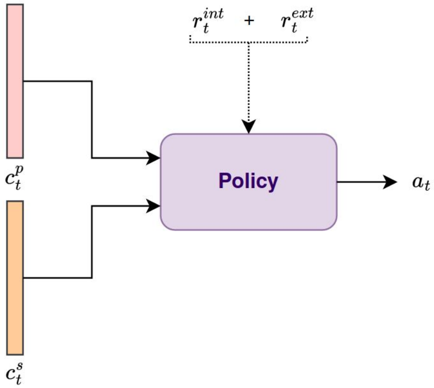
Options for training:

- From reward
- From message quality
- From mutual information with inputs

I. Language-Augmented MARL

Policy

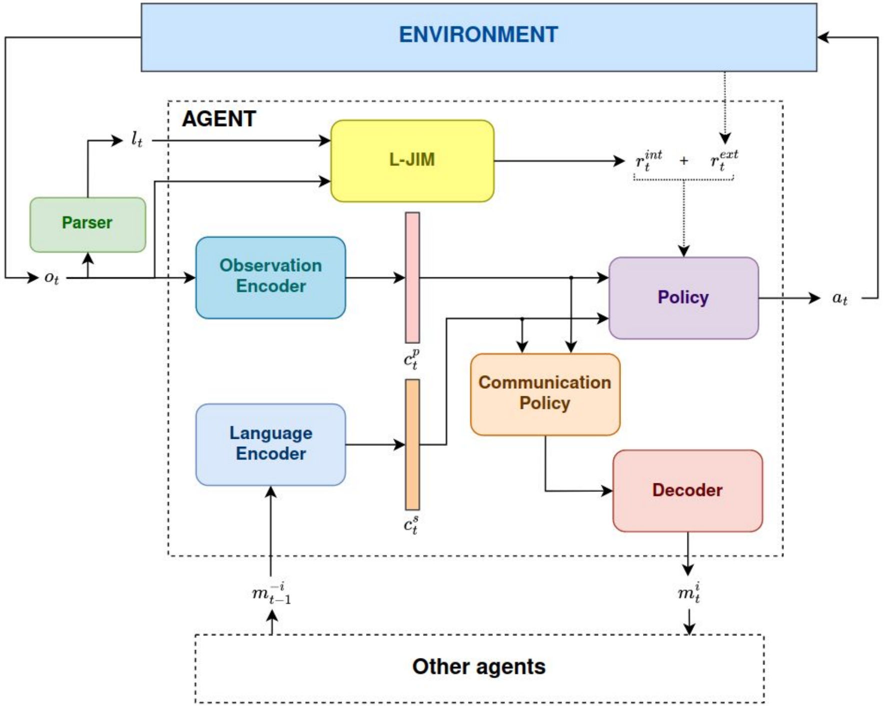




Trained with any MADRL algorithm.

I. Language-Augmented MARL

Architecture



Algorithm:

1. Declare for each agent (some or all of these networks may share parameters across agents):
    - the observation encoder Ω,
    - the language encoder Λ,
    - the decoder Δ,
    - the policy Π potentially with some value function,
    - the communication policy C,
    - L-JIM
2. Initialise a replay buffer
3. For $t = 0, 1, \ldots, N_{frames}$, do:
    1. Get observations o_t from the environment
    2. Parse o_t to get the language description l_t
    3. Encode the observation into the internal context embedding: $c_t^p = \Omega(o_t)$
    4. Encode the messages received from other agents (concatenated) into the social context embedding: $c_t^s = \Lambda(m_{t-1}^{-i})$
    5. Generate a message to send to other agents: $m_t^i = \Delta(C(c_t^p, c_t^s))$
    6. Generate action to perform: $a_t = \Pi(c_t^p, c_t^s)$
    7. Compute intrinsic reward: $r_t^{int} = \text{L-JIM}(o_t, l_t)$
    8. Perform action, get new observation o_{t+1} and external reward r_t^{ext} from the environment
    9. Store experience (o_t, a_t, r_t, o_{t+1}), with $r_t = r_t^{int} + r_t^{ext}$, in the replay buffer (add l_t?)
    10. If $t\%100 = 0$:
        1. Sample batch of experience from replay buffer
        2. Train decoder with observation captioning
        3. Train language encoder with CLIP
        4. Train communication policy
        5. Train policy
        6. Train L-JIM

II. Joint Intrinsic Motivation

Context

- MADRL algorithms struggle with sparse reward
- Relative overgeneralization

	A	B	C
A	10	-5	-5
B	-5	7	7
C	-5	7	7

Figure: Social dilemma game where relative overgeneralization occurs.

Context

- MADRL algorithms struggle with sparse reward
- Relative overgeneralization

- In single-agent RL, intrinsic rewards are used to incite active exploration of the environment

	A	B	C
A	10	-5	-5
B	-5	7	7
C	-5	7	7

Figure: Social dilemma game where relative overgeneralization occurs.

Context

- MADRL algorithms struggle with sparse reward
- Relative overgeneralization

- In single-agent RL, intrinsic rewards are used to incite active exploration of the environment

	A	B	C
A	10	-5	-5
B	-5	7	7
C	-5	7	7

Figure: Social dilemma game where relative overgeneralization occurs.

Our solution

- Reward agents for finding new joint behaviors with Joint Intrinsic Motivation

Related works

Multi-agent Deep Reinforcement learning

- Centralized Training with Decentralized Execution (CTDE)
  - MA-DDPG (Lowe et al., *Multi-agent actor-critic for mixed cooperative-competitive environments*, NeurIPS, 2017)
  - MA-PPO (Yu et al., *The surprising effectiveness of ppo in cooperative, multi-agent games*, 2021)
- Credit assignment
  - COMA (Foerster et al., *Counterfacutal multi-agent policy gradients*, AAAI, 2018)
- Value factorisation methods
  - VDN (Sunehag et al., *Value-decomposition networks for cooperative multi-agent learning based on team reward*, AAMAS, 2018)
  - **QMIX** (Rashid et al., *QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning*, ICML, 2021)
  - Qtran (Son et al., *Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning*, ICML, 2019)
  - MAVEN (Mahajan et al., *Multi-agent variational exploration*, NeurIPS, 2019)
  - W-QMIX (Rashid et al., *Weighted qmix: Expanding monotonic value function factorisation for deep multiagent reinforcement learning*, NeurIPS, 2020)

Single-agent intrinsic rewards

- ICM (Pathak et al., *Curiosity-driven exploration by self-supervised prediction*, 2017)
- RND (Burda et al., *Exploration by random network distillation*, ICLR, 2019)
- RIDE (Raileanu and Rocktäschel, *Ride: Rewarding impact-driven exploration for procedurally-generated environments*, ICLR, 2020)
- NGU (Badia et al., Never give up: Learning directed exploration strategies, ICLR, 2020)
- AGAC (Flet-Berliac et al., *Adversarially guided actor-critic*, ICLR, 2021)
- **NoveID** (Zhang et al., *Noveld: A simple yet effective exploration criterion*, NeurIPS, 2021)
- **E3B** (Henaff et al., *Exploration via elliptical episodic bonuses*, NeurIPS, 2022)

Multi-agent intrinsic rewards

- Social inluence
  - Jaques et al.,*Social influence as intrinsic motivation for multi-agent deep reinforcement learning*, ICML, 2019
  - Wang et al., *Influence-based multi-agent exploration*, ICLR, 2020
- Alignment
  - ELIGN (Ma et al., *ELIGN: Expectation alignment as a multi-agent intrinsic reward*, NeurIPS, 2022)
- Credit assignment
  - LIIR (Du et al., *Liir: Learning individual intrinsic reward in multi-agent reinforcement learning*, NeurIPS, 2019)
- Coordinated exploration
  - Multi-Explore (Iqbal and Sha, *Coordinated exploration via intrinsic rewards for multi-agent reinforcement learning*, 2019)

II. Joint Intrinsic Motivation

Method: Double-timescale intrinsic reward

$$r_t^{int}(s_t, a_t, s_{t+1}) = N_l(s_t, s_{t+1}) \times \sqrt{2b(s_{t+1})}$$

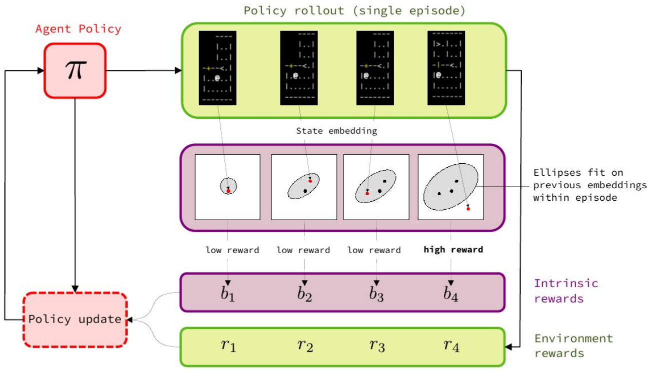
- Life-long exploration criterion: NovelD => Explore unknown regions of the environment

$$N_l(s_t, s_{t+1}) = \max[RND(s_{t+1}) - \alpha RND(s_t), 0]$$

- Episodic exploration criterion: E3B => Experience more diverse trajectories

$$b(s_t) = \psi(s_t)^\top C_{t-1}^{-1} \psi(s_t),$$

$$C_{t-1} = \sum_{i=1}^{t-1} \psi(s_i)\psi(s_i)^\top + \lambda I$$



Method: Joint Intrinsic Motivation

$$r_t = r_t^e + \beta r_t^{JIM}$$

Use joint observations for computing the reward => Search for novelty in the joint-observation space

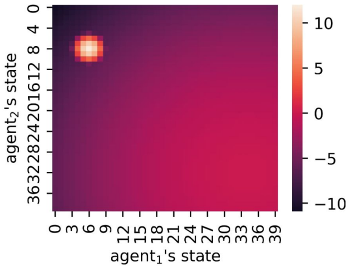
$$r_t^{JIM}(\mathbf{o}_t, \mathbf{a}_t, \mathbf{o}_{t+1}) = N_l(\mathbf{o}_t, \mathbf{o}_{t+1}) \times \sqrt{2b(\mathbf{o}_{t+1})}$$

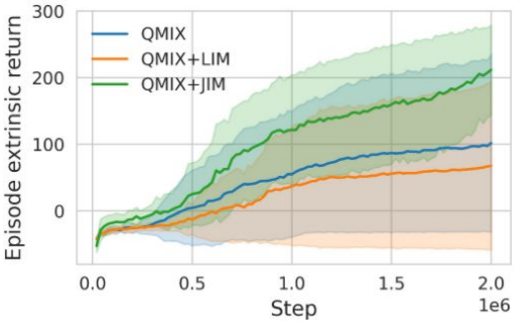
Can be used with any MADRL algorithm that uses CTDE

II. Joint Intrinsic Motivation

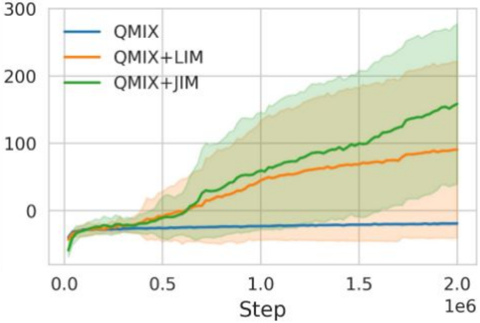
Experiments: Addressing relative overgeneralization (rel_overgen)

	A	B	C
A	10	-5	-5
B	-5	7	7
C	-5	7	7

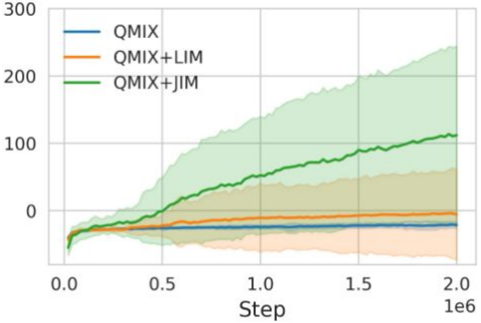




(a) easy ($\delta = 30$)



(b) hard ($\delta = 40$)



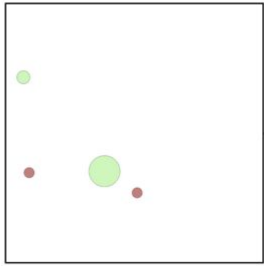
(c) very hard ($\delta = 50$)

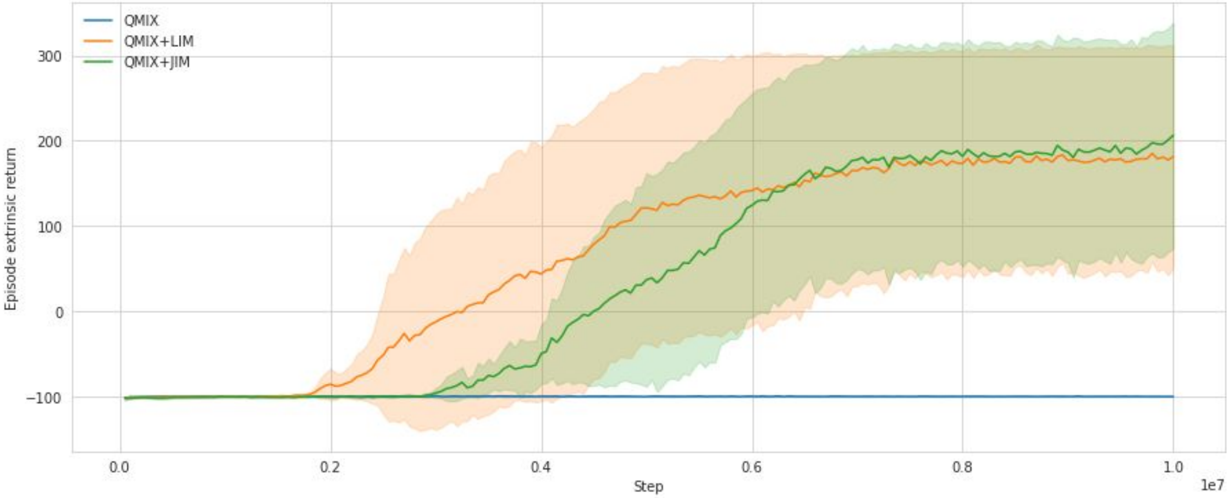
II. Joint Intrinsic Motivation

Experiements: Cooperative push

Task: Push an object on a landmark.

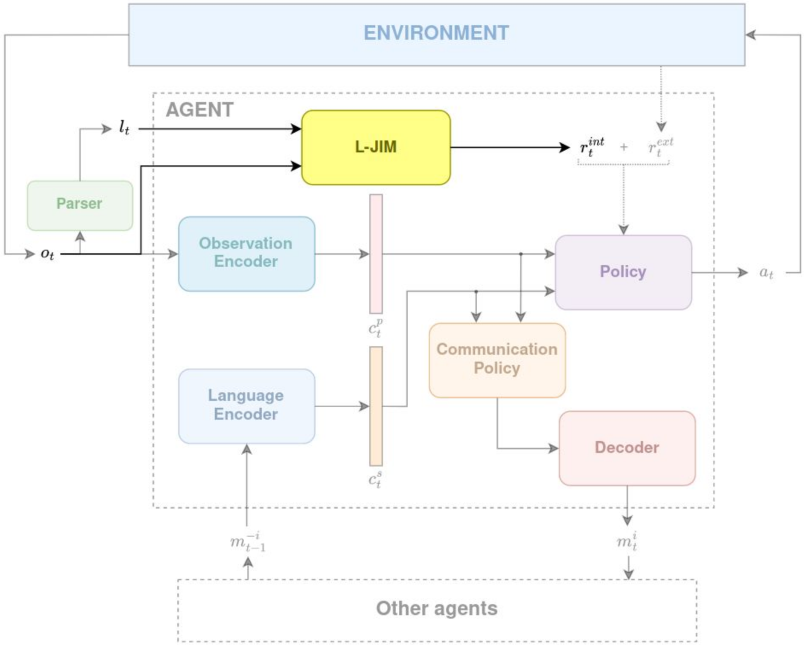
Reward: +400 if object on landmark, -1 every step





II. Joint Intrinsic Motivation

In the Language-Augmented MARL architecture



II. Joint Intrinsic Motivation

Next steps

Experiments

- Find push scenario that validates results on rel_overgen
- Train for more steps on rel_overgen
- Train more runs for each scenario

Paper

- Finish writing
- Submit to ECAI (next April)

Language-Augmented MARL

- Adapt JIM for language

III. Next year

Planning

- Publish JIM (now->April)
- Develop LA-MADRL approach (now->September)
  - Add JIM
  - Communication policy
  - Publish
- Writing memoire (June->November)
- Prepare defense (December)
- Thesis defense (end of December)

IV. Training plan

- Mobile robotics course, master 2, Sorbonne Université (36 hours)
- Scientific writing formation (12 hours)
- Teaching programming for 1st year ECE Paris students (100hours/year)

Need 13 more hours of non-technical formations

Thank you for you attention !

Questions ?