Learning to use a pre-defined language: Training a Communication Policy

Maxime Toquebiau[1,2], Nicolas Bredeche[2], Faïz Ben Amar[2], Jae Yun Jun Kim[1]

[1]ECE Paris
[2]ISIR, Sorbonne Universités

September 13th, 2023

Table of Content

- JIM Paper
- Communicating with language

Joint Intrinsic Motivation

Re-Submission

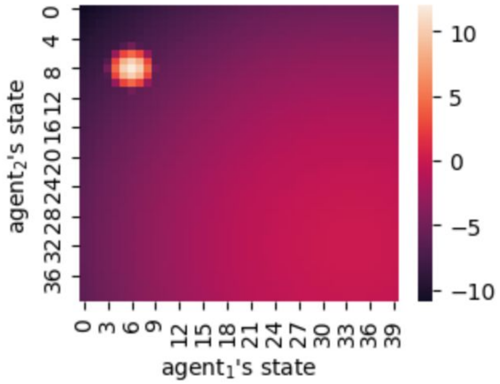
Goal: re-submission to AAMAS, with:

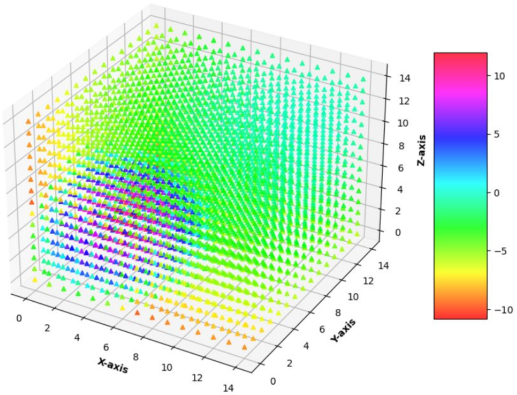
- Scenario with 4 agents:

→ `rel_overgen`, playing with hyperparameters:

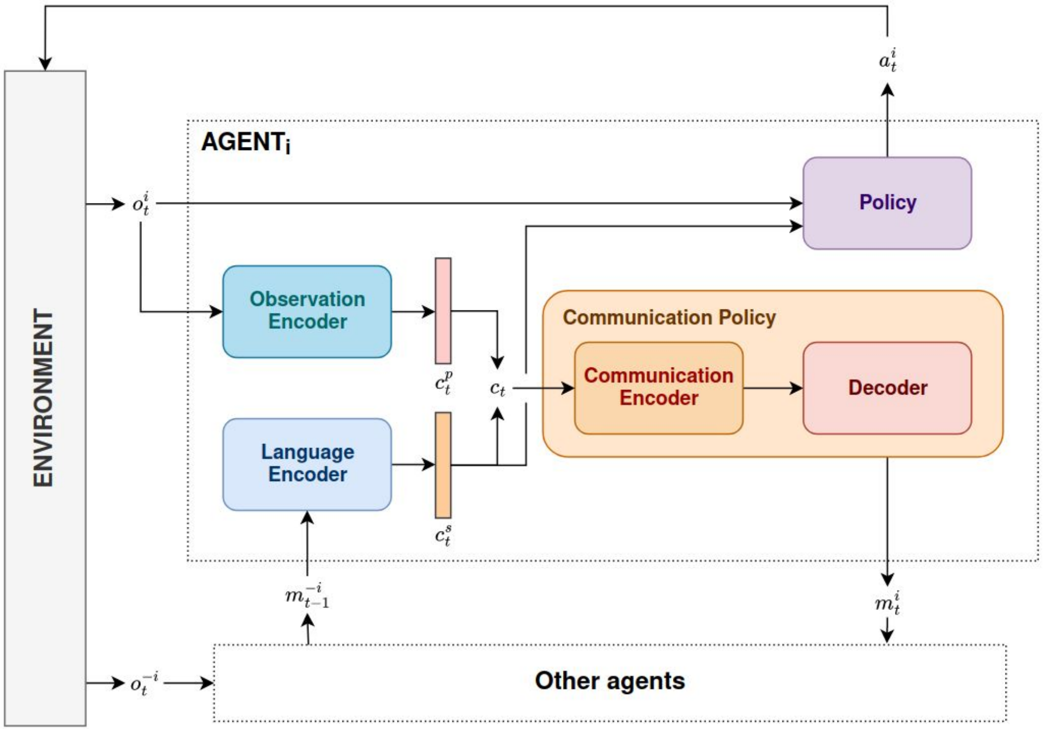
  - state dimension,
  - ε-greedy exploration probability,
  - size of reward spike.

- Extended explanation of scalability in paper

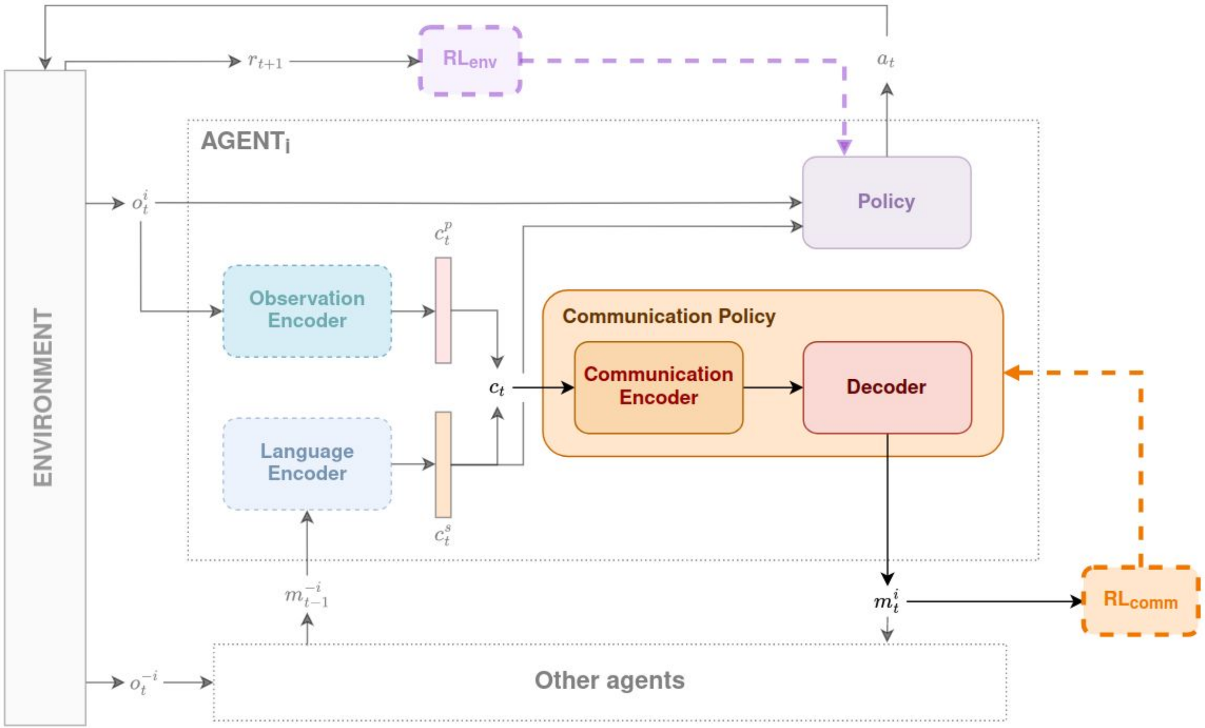




Communicating with language

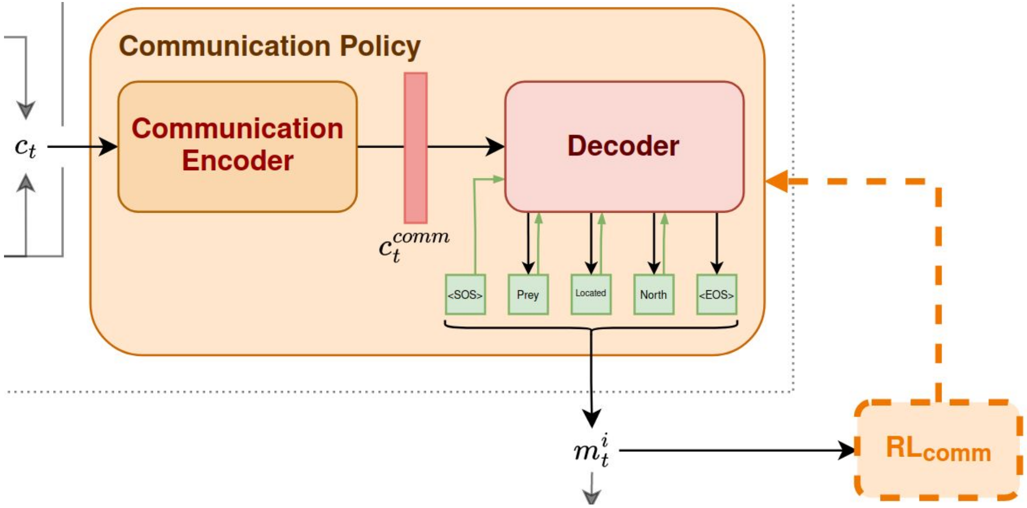


Communicating with language



Communication Policy

Training policy

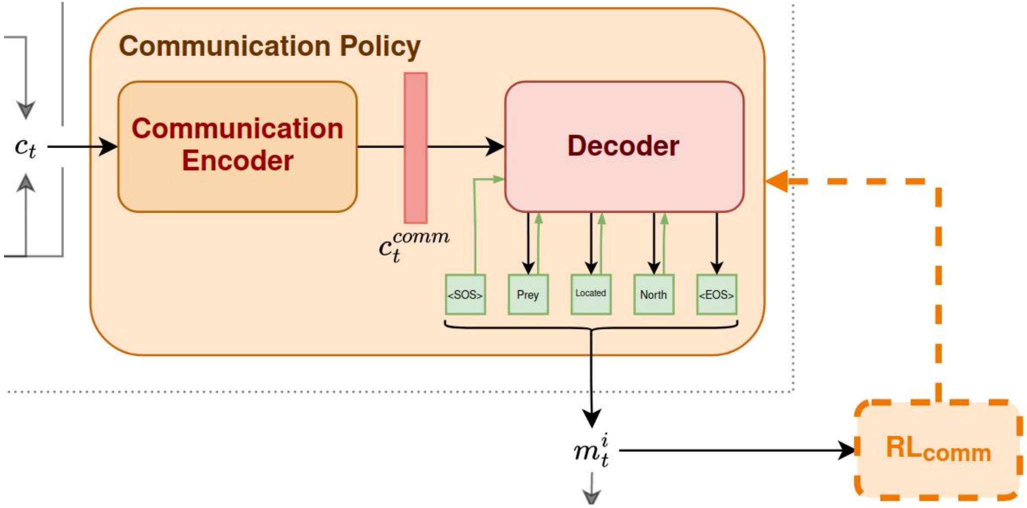


Training communication with PPO:

- **Task:** Generating messages (sequences of tokens)
- **States:** previous token
- **Actions:** next token

Communication Policy

Communication Encoder



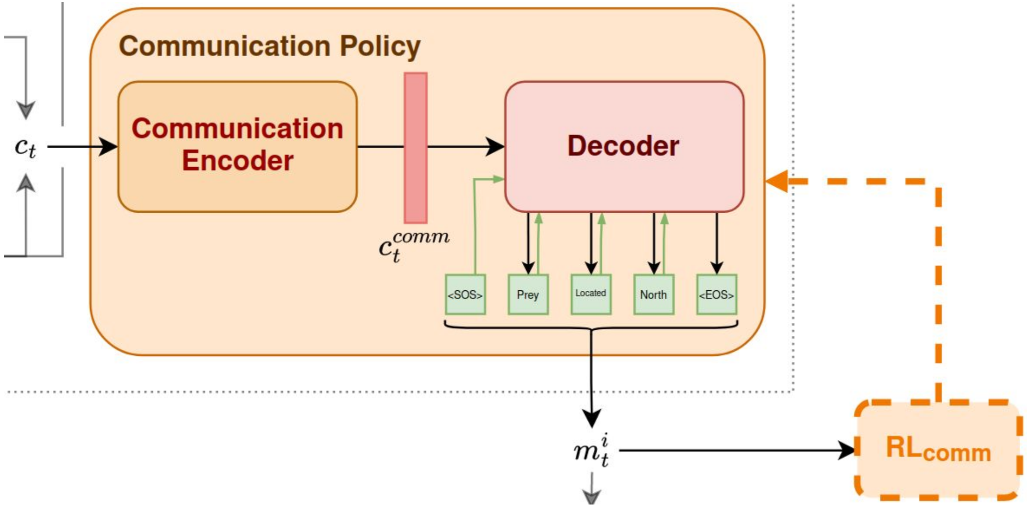
Communication Encoder:

- **MLP**
  - single step as input
  - stack of steps as input
- **RNN**
- **Hierarchical** (Jaques2019, Saleh2020)

[1] Jaques et al., *Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog*, 2019.
[2] Saleh et al., *Hierarchical Reinforcement Learning for Open-Domain Dialog*, AAAI Conf. on AI 2020.

Communication Policy

Evaluating communication quality



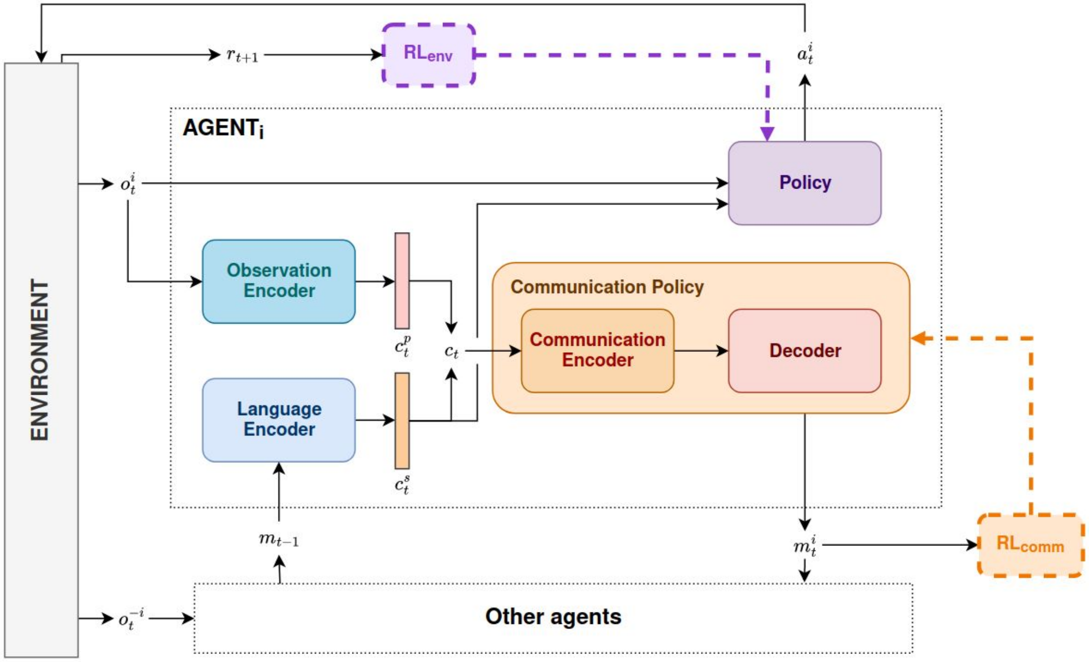
Communication Evaluation:

- **Task performance:** Reward from environment
- **Language drifting:** Penalty for diverging too much from pre-trained decoder (Ouyang2022)
- **Efficient communication:** Penalty for each token generated
- **Sharing valuable information:** Reward for adding information to a shared-memory
- **Impacting other agents:** maximizing information gain, wonderful life…
- …

[1] Ouyang et al., *Training language models to follow instructions with human feedback*, 2022.

Communicating with language

Training issue



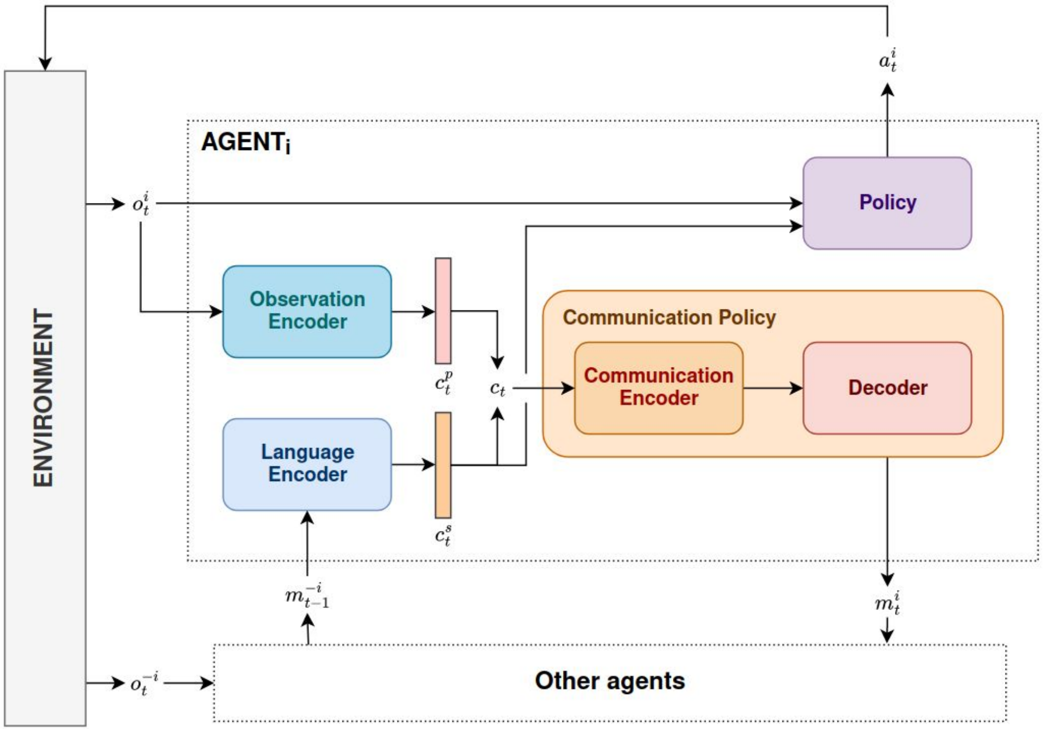
Two Parallel Training Loops:

Act Loop: RL from environment reward *trained after each episode*

vs.

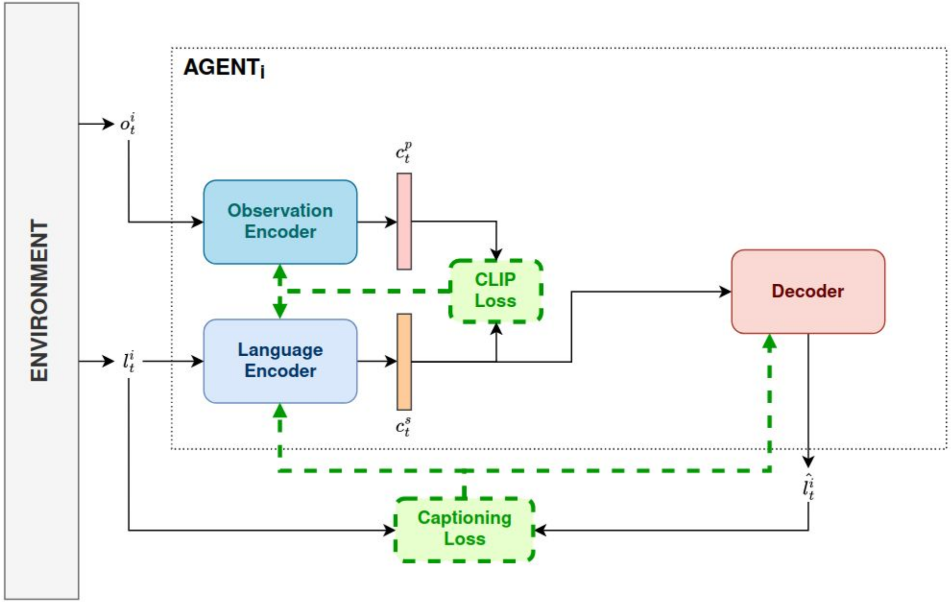
Communication Loop: RL from communication quality *trained after each step*

Communicating with language



Training Process

Phase 1: Learning to ground and generate language



CLIP Loss

With the **Observation Encoder** $\omega : \mathbb{R}^N \to \mathbb{R}^M$,
and the **Language Encoder** $\lambda : \mathbb{R}^{L \times V} \to \mathbb{R}^M$,

the grounding objective is:

$$J(\theta_\omega, \theta_\lambda) = max[cosim(\omega(o_k), \lambda(l_k))]$$

+

Captioning Loss

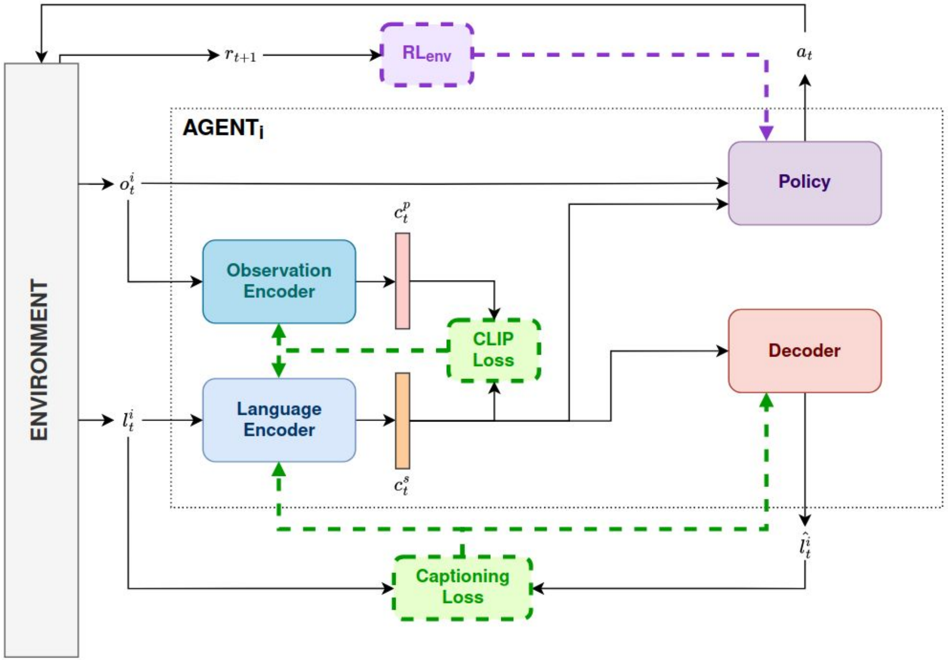
With the **Language Encoder** $\lambda : \mathbb{R}^{L \times V} \to \mathbb{R}^M$,
and the **Decoder** $\delta : \mathbb{R}^M \to \mathbb{R}^{L \times V}$,

the captioning objective is:

$$J(\theta_\lambda, \theta_\delta) = min\left[\frac{1}{N}\sum_{i=0}^{N}(\hat{l}_i - l_i)^2\right]$$

Training Process

Phase 2: Learning a working policy with “perfect messages”



CLIP Loss

+

Captioning Loss

+

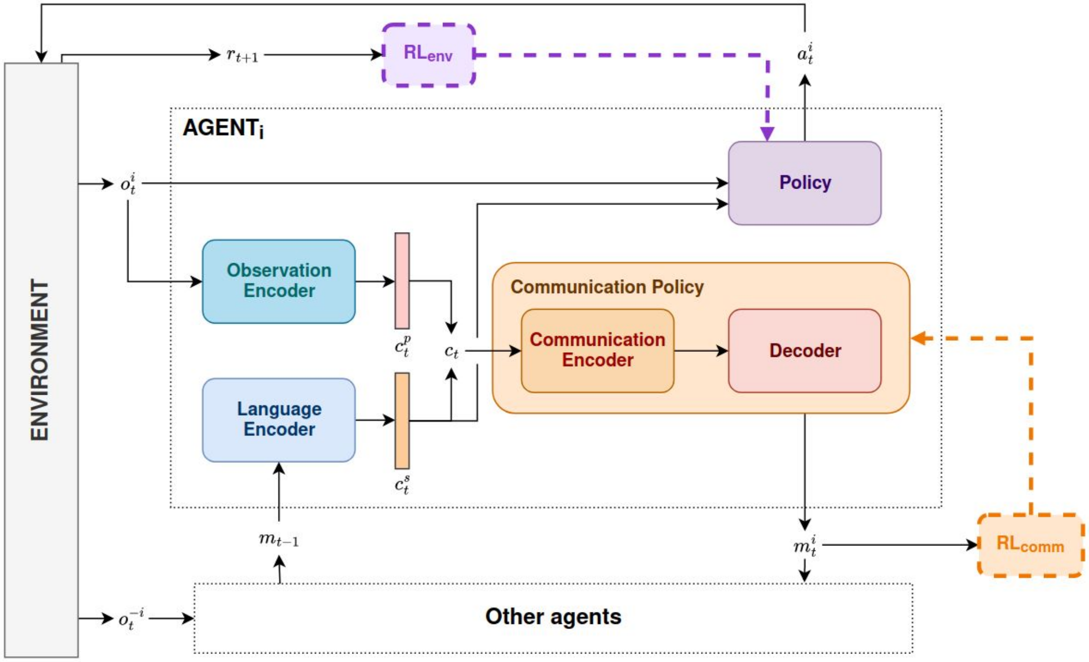
RL from Environment rewards

Training with any reinforcement learning algorithm using rewards from the environment.

Most likely **MAPPO**, so with, for each agent i, an **Actor** $\pi^i_{\phi_{act}}(o^i_t) = a^i_t$ and a **Critic** $V^i_{\phi_{crit}}(\mathbf{o}_t)$, learning to maximise rewards r_t from the environment by optimising the PPO objective

Training Process

Phase 3: Learning the communication policy



Fixed parameters

RL$_{\text{env}}$

+

RL from Communication quality

We train a **Communication Encoder** $\mathbb{C} : \mathbb{R}^{2M} \to \mathbb{R}^{M}$ to learn to choose which information to share, and we fine-tune the **Decoder** δ (pre-trained on captioning) to generate useful messages.

The reward for communication quality can be defined as,

$$r_t^{comm} = r_t - \beta \log \left(\frac{\delta_{FT}}{\delta_{PT}} \right) + \dots$$

with δ_{FT} the current fine-tuned version of the decoder and δ_{PT} the pre-trained version of the decoder with fixed parameters. We use PPO for learning the communication policy.

Next steps

- JIM:
    - Get working runs with 4 agents
    - Finish paper

- Communication policy
    - Dev communication evaluation
    - Move to BabyAI environment
    - Experiment with Communication Policy
    - Write paper

AAMAS 2024 deadline: October 9th

Thank you for you attention !

Questions ?