Communication-based Cooperative Tasks: how the Language Expressiveness affects Reinforcement Learning

Jacopo Talamini, Eric Medvet, Alberto Bartoli

DIA - University of Trieste

09/04/2019
Introduction

Multi-agent systems (MAS)
Models in which many artificial agents act in the same environment with common and/or conflicting goals. Many applications:

- IoT devices (domotics, 5G technology)
- Autonomous vehicles
- Smart cities
- Industrial robots
- ...

[Images of autonomous vehicles, robots, smart cities, and internet of things devices]
Communication-based MAS

One way to allow agents to interact is by introducing communication.

How to communicate?
  ▶ Pre-defined rules for all the agents
  ▶ No pre-defined rules: communication is learned
  ▶ ...

Which type of language?
  ▶ Natural language
  ▶ Symbol-based language
  ▶ ...
Communication-based MAS (II)

Hypothesis

▶ Agents learn communication
▶ Symbol-based language

Research questions

▶ Which design choices are best for a given scenario?
▶ To which extent language design affect task performance?
▶ Which are the limits of language design in terms of learning?
Research plan

1. Design a scenario for communication-based MAS
2. Define a strategy for learning communication
3. Experimentally investigated research questions
Markov Decision Process (MDP)

Graph in which state transitions depend only on the current state and action taken:

\[
p(s', r | s, a) = \Pr\{s_{t+1} = s', r_{t+1} = r | s_t = s, a_t = a\} \tag{1}
\]

A MDP is described by \((S, P, A, \pi, R)\)
Markov Game

We consider a multi-agent partially observed MDP (Markov Game)

A Markov Game of $n$ agents is described by $(S, \phi, \rho, O, A, \Omega, \Pi, R)$

- $S$ is the set of possible states of the game
- $\phi$ is the stochastic state transition function
- $\rho$ is the stochastic initial state function
- $O = (O_1, \ldots, O_n)$ is the set of agent observations
- $A = (A_1, \ldots, A_n)$ is the set of agent actions
- $\Omega = (\omega_1, \ldots, \omega_n)$ is the set of agent observation functions
- $\Pi = (\pi_1, \ldots, \pi_n)$ is the set of agent stochastic policies
- $R = (r_1, \ldots, r_n)$ is the set of agent reward functions
Cooperative Unknown Target (CUT) game

Scenario
We consider a cooperative MAS, in which:

▶ The environment is a 2D world with agents and targets
▶ Agents can move and communicate
▶ Targets are deployed at random

Each agent observes:

▶ Location of every target
▶ Another agent-target pairing (\(a\) knows the target of \(b\))
▶ Words spoken by the other agents

Objective

▶ Minimize distance from one specific **unknown** target
▶ Guide other agents to their target using communication
Cooperative Unknown Target (CUT) game (II)

Learning problem

We frame this task as a learning problem that we tackle using Reinforcement Learning (RL)

Agents learn how to:

▶ Emit words, given target-agent pairings
▶ Move to a specific target, given targets position and words spoken by the other agents
CUT as Markov Game
Each agent in CUT is mapped into a Markov Game as 2 agents:

- **Movement agent**
  - $O^{\text{move}} = \mathbb{R}^2 \times \ldots \times \mathbb{R}^2 \times \mathcal{W} \cup \emptyset \times \ldots \times \mathcal{W} \cup \emptyset$
  - $A^{\text{move}} = \{\uparrow, \rightarrow, \downarrow, \leftarrow, \emptyset\}$
  - $r_i^{\text{move}}(s) = \sum_{j=1}^{n_r} -d_1(\bar{x}_{r,i}, \bar{x}_{t,\tau_{i,j}}) \mathbb{1}_i(s, j)$

- **Communication agent**
  - $O^{\text{comm}} = \{1, \ldots, n_r\} \times \{1, \ldots, n_t\}$
  - $A^{\text{comm}} = \mathcal{W} \cup \emptyset$
  - $r_i^{\text{comm}}(s) = -d_1(\bar{x}_{r,\tau_r,i}, \bar{x}_{t,\tau_{t,i}})$
**RL strategy**

- Individual policies: every agent may learn to speak in its own language
- Issue: environment is **non-stationary** during the learning process
- Solution: Multi-Agent Deep Deterministic Policy Gradient (MADDPG):
  - Centralized Critic, decentralized actors
  - Actors learn using estimation of other policies
Language expressiveness

Given this learning scenario, with \( n_r \) agents and \( n_t \) targets, and language vocabulary \( W \)
We define the language expressiveness \( e \) as:

\[
e = \frac{|W|}{n_r n_t}
\]  \hspace{1cm} (2)

We investigate the impact of \( e \) on learning in terms of:

- Effectiveness
- Efficiency
Experimental evaluation

Experiment settings

- Simultaneous policy learning of $\pi^{\text{move}}$ and $\pi^{\text{comm}}$ independently
- Different combinations of $(n_r, n_t, |W|)$ with $e \in [0, 4]$:
  - $(n_r, n_t) \in \{2, 3, 4\} \times \{2, 3, 4\}
  - $|W| \in \{0, \ldots, 4n_r n_t\}$
- Baselines:
  - No-communication
  - Optimal communication ($\pi^{\text{comm}}_{\text{Opt}}$)

Training settings

For each combination of $(n_r, n_t, |W|)$:

- 20000 training episodes ($R_{\text{learn}}$)
- 100 validation episodes ($R_{\text{val}}$)
Effectiveness: overview

Standardized validation reward $R_{val}$

Highest validation reward when $e \sim 1$, i.e., when $|W| \sim n_r n_t$
Effectiveness: problem size

\( n_r, n_t \) small:
- Constant validation reward

\( n_r, n_t \) big:
- Degrading behavior
- Higher gap \( \pi_{\text{Opt}}^{\text{comm}} \) baseline-learned policies
Effectiveness: variability

Significant variability in the outcome of policy learning, more so when expressiveness is large.
Efficiency

We consider a different setting for each plot:

- \( n_r = n_t = 2 \), high learning efficiency, \( \pi_{\text{NoComm}}^{\text{comm}} \) and the policy with \( e = 0.25 \) sharply separated from the others
- \( n_r = n_t = 3 \), the baseline policies exhibit more efficiency than the others
- \( n_r = n_t = 4 \), the baseline \( \pi_{\text{Opt}}^{\text{comm}} \) is not efficient, yet reaches the higher \( R_{\text{val}} \)

\[
\begin{align*}
\text{nr} = 2, \ n_t = 2 & \\
\text{nr} = 3, \ n_t = 3 & \\
\text{nr} = 4, \ n_t = 4 & \\
\end{align*}
\]
Conclusions

- We trained agents to play the CUT game
- We investigated the impact of expressiveness on learning performances in terms of:
  - Effectiveness
  - Efficiency