Effects of Supervision, Population Size, and Self-Play on Multi-Agent Reinforcement Learning to Communicate

Marina Dubova (dubova_marina) [1] **Arseny Moskvichev** (arseny_mo) [2]

[1]Indiana University [2]UC Irvine

Modeling Language Evolution

- Language evolution is driven by individual and population-level factors (e.g. **learning biases** and **social structure**).
- Modeling is essential to uncover the key properties of language evolution. Many models have been proposed, but most of them focus on formalizing the adaptation and transmission processes on only one of the levels: **individual** (e.g. robotic models) or **populational** (e.g. genetic and iterative learning models).
- Connectionist multi-agent modeling framework offers a good compromise: the models can be scaled to represent long-term learning in relatively big populations without sacrificing the individual learning dynamics.

Multi-Agent Reinforcement Learning as a Framework for Language Evolution Research: Challenges and Proposed Solutions

Challenges of decentralized deep learning models of language evolution:

- **Instability**: optimization problem for each agent is non-stationary and non-Markov
- **Sparsity of rewards**
- **Enormous state sizes**

Solutions used by the deep learning community:

- **Centralized optimization** (e.g. training a shared controller for all the agents, adopting inter-agent weight-sharing, etc)
- Restricting models to purely **supervised learning**
- Using **small population sizes** (2-4 agents)

Research goal

Humans are individually flexible and independent in their learning, but they successfully converge on shared communicative systems. This may be largely due to the regularizing properties of the environment in which they operate. **We investigated the effects of naturalistic interventions on decentralized multi-agent communication learning outcomes**.

Multi-Agent Coordination Game

On every step, two agents are selected to play the game, and the speaker and listener roles for the episode are assigned. The speaker produces a message and an action, the listener receives a message and produces an action. If the actions match, both agents receive reward (Figure 1). To avoid trivial solutions, we penalize overusing any specific action.

Each agent is represented with a simple feed-forward neural network. The networks were trained using a vanilla deep Q-learning algorithm.

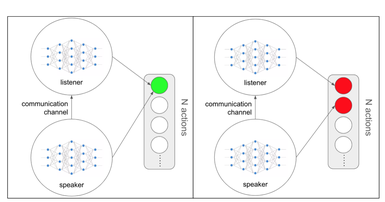

Figure 1:Coordination Game. Left: successful coordination (reward), right: unsuccessful coordination (no reward)

Interventions

- Bottom-driven **supervision** (from 0 to 0.9). A fraction of trials is used to provide a supervising signal (i.e. the action that the other agent used is treated as a "rewarding" answer).
- **Population size** (from 2 to 6).
- **Self-play** (can be interpreted as ``inner speech", present / absent).

Metrics

1. **Speaking Consistency and Listening Consistency**. Normalized mutual information between the distribution over messages and actions that an agent defines.

$$S/LC = \sum_{a \in A_l} \sum_{m \in A_c} p_{a,m}(a,m) log \frac{p_{a,m}(a,m)}{p_a(a) p_m(m)} / Z \quad (1)$$

2. **Between-agent signal-action mapping divergence (homogeneity)**. Average Jensen-Shannon pairwise divergence between distributions of agents' actions following a specific signal (averaged over all signals and pairs of agents).

$$\sum_{m \in A_c} JSD(p_{a_1|m}, p_{a_2|m}) / |A_c| \quad (2)$$

3. **Within-agent signal-action mapping divergence**. Average Jensen-Shannon divergence between the distributions over agent's actions when the agent plays speaker and listener roles (conditioned on receiving or sending a specific message).
4. **Talking divergence**. Average pairwise Jensen-Shannon divergence of marginal signaling distributions of different agents.
5. **Behavioral Predictability**. Jensen-Shannon divergences between marginal distributions of agent's actions and messages and the uniform distribution.

Experiment 1: self-play and supervision rate

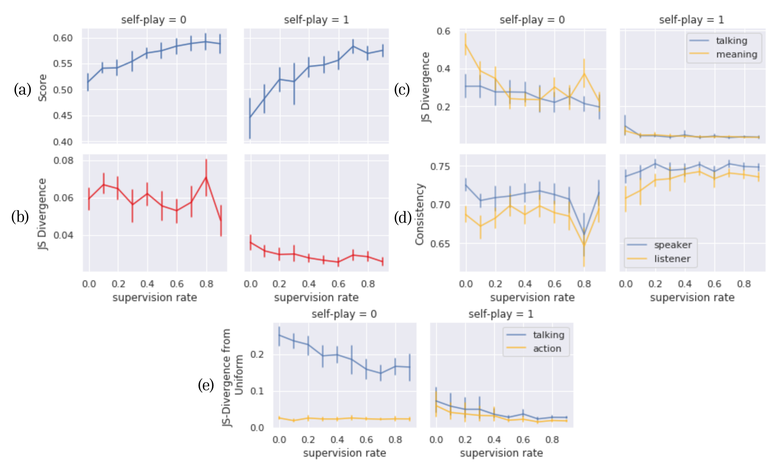

Figure 2:The Communication Analysis Metrics with 95% Confidence Intervals for the Experiment 1. a) Average rewards b) Between-agent signal-action mapping divergence c) Talking divergence (blue) and Within-agent signal-action mapping divergence (yellow) d) Average speaking (blue) and listening (yellow) consistencies e) Average predictability of agents' signals (blue) and actions (yellow).

Experiment 2: population size and supervision rate

Figure 3:The Communication Analysis Metrics with 95% Confidence Intervals for the Experiment 1. a) Average rewards b) Between-agent signal-action mapping divergence c) Talking divergence (blue) and Within-agent signal-action mapping divergence (yellow) d) Average speaking (blue) and listening (yellow) consistencies e) Average predictability of agents' signals (blue) and actions (yellow).

Results

- Naive reinforcement feedback is not sufficient to ensure that independently adapting agents converge on effective communication systems
- Independent learning in a group of two agents results in internally and externally asymmetric communication
- Both bigger agent population and self-play learning introduce a pressure to converge on **shared** and **symmetric** communication systems
- Adding supervising feedback positively affected agents' communication performance across all metrics, but it was not as efficient in solving the asymmetry issues as introducing self-play or increasing the population size
- Reward scores do not provide enough information on the learning results. A **comprehensive set of metrics is required** to reliably interpret the results of multi-agent communication learning

Acknowledgements

Authors thank Robert Goldstone, Denizhan Pak, Jack Avery, Mahi Luthra, Brian Dahlberg, all the attendees of the PCL lab meetings, and two anonymous reviewers for their valuable feedback on this project. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. https://kb.iu.edu/d/anwt#carbonate