Learning Policy Representations in Multiagent Systems

Aditya Grover[1], Maruan Al-Shedivat[2], Jayesh K. Gupta[1], Yura Burda[3], Harrison Edwards[3]
[1]Stanford University [2]Carnegie Mellon University [3]OpenAI

ABSTRACT

Modeling agent behavior is central to understanding the emergence of complex phenomena in multiagent systems (MAS). We propose a general learning framework for modeling agent behavior in any multiagent system using only a handful of interaction data.

Our framework casts agent modeling as a representation learning problem. Consequently, we construct a novel objective inspired by imitation learning and agent identification and design an algorithm for unsupervised learning of representations of agent policies.

MOTIVATION

Intelligent agents participate in diverse, complex interactions with each other in competitive and/or cooperative environments. Our framework is motivated by the following questions:

- **Learning:** How can we learn representations of agent policies via interaction episodes?
- **Inference:** For which tasks are learned policy representations useful?
- **Evaluation:** What constitutes generalization in a multiagent system?

LEARNING OBJECTIVE

We propose two desiderata for the learned representations in multiagent systems.

1. **Generative.** The representation should be useful for simulating the agent's policy.
   → **Conditional Imitation Learning**

2. **Discriminative.** The representation should be able to distinguish the agent's policy with the policies of other agents.
   → **Agent Identification using Triplet Loss**

ALGORITHM

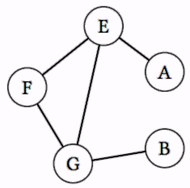
Algorithm 1 Learn Policy Embedding Function (f_θ)

input $\{E_i\}_{i=1}^n$ – interaction episodes, λ – hyperparameter.
Initialize θ and ϕ
for $i = 1, 2, \ldots, n$ **do**
  Sample a positive episode $p_e \leftarrow e_+ \sim E_i$
  Sample a reference episode $r_e \leftarrow e_* \sim E_i \backslash e_+$
  Compute `Im_loss` $\leftarrow -\sum_{\langle o,a \rangle} \log \pi_{\theta,\phi}(a|o, e_*)$
  for $j = 1, 2, \ldots, n$ **do**
    if $j \neq i$ **then**
      Sample a negative episode $n_e \leftarrow e_- \sim E_j$
      Compute `Id_loss` $\leftarrow d_\theta(e_+, e_-, e_*)$
      Set `Loss` $\leftarrow$ `Im_loss` $+ \lambda \cdot$ `Id_loss`
      Update θ and ϕ to minimize `Loss`
    end if
  end for
end for
output θ

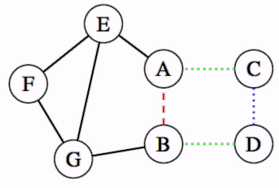
GENERALIZATION IN MAS

Agent-Interaction Graphs (AIG)

Nodes → Agent policies
Edges → Observed Interaction Episodes



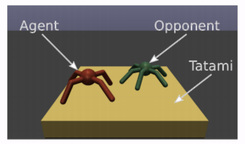
Train AIG



Train + Test AIG

Weak generalization: New edges (AB)
Strong generalization: New nodes (C, D)

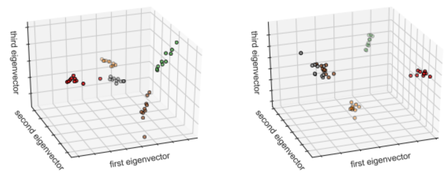
EXPERIMENTS



RoboSumo

25 PPO agents:
MLP (2 hidden layers)

Embedding function:
MLP (3 hidden layers)



Projection of Emb-Hyb embeddings on first 3 principal components for weak (left) and strong (right) generalization. Color denotes agent policy.

POLICY OPTIMIZATION WITH LEARNED EMBEDDINGS

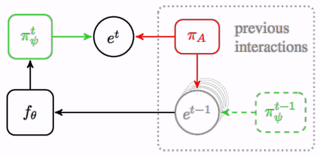


Figure 2: Illustration of the proposed model for optimizing a policy π_ψ that conditions on an embedding of the opponent policy π_A. At time t, the pre-trained representation function f_θ computes the opponent embedding based on a past interaction e^{t-1}. We optimize π_ψ to maximize the expected rewards in its current interactions e^t with the opponent.

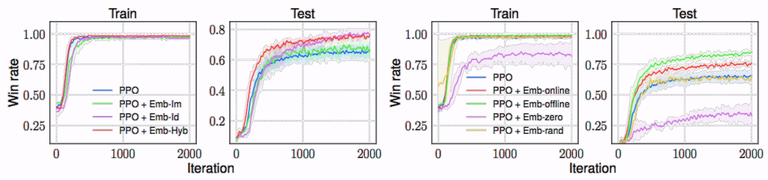


Figure 4: Average win-rates of the newly trained agents against 5 training agent and 5 testing agents. The left two charts compare baseline with policies that make use of `Emb-Im`, `Emb-Id`, and `Emb-Hyb` (all computed online). The right two charts compare different embeddings used at evaluation time (all embedding-conditioned policies use `Emb-Hyb`). At each iteration, win-rates were computed based on 50 1-on-1 games. Each agent was trained 3 times, each time from a different random initialization. Shaded regions correspond to 95% CI.

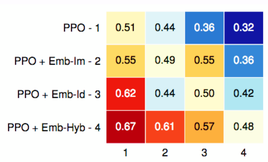


Figure 6: Win-rates for agents specified in each row at computed at iteration 1000.

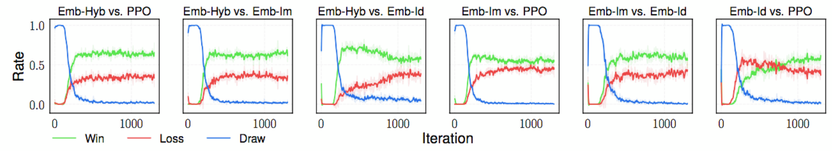
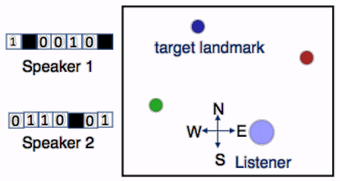


Figure 5: Win, loss, and draw rates plotted for the first agent in each pair. Each pair of agents was evaluated after each training iteration on 50 1-on-1 games; curves are based on 5 evaluation runs. Shaded regions correspond to 95% CI.

See paper for experiments on Cooperative Speaker-Listener Particle Environment!



REFERENCES

1. Al-Shedivat, M., Bansal, T., Burda, Y., Sutskever, I., Mordatch, I., and Abbeel, P. Continuous adaptation via metalearning in nonstationary and competitive environments. In ICLR, 2018.
2. Grover, A., Al-Shedivat, M., Gupta, J. K., Burda, Y., and Edwards, H. Evaluating generalization in multiagent systems using agent-interaction graphs. In AAMAS, 2018.
3. Wang, Z., Merel, J., Reed, S., Wayne, G., de Freitas, N., and Heess, N. Robust imitation of diverse behaviors. In NIPS, 2017