Multi-agent Reinforcement Learning

Achieve Cooperation and Competition with Implicit Intent Inference

Hui Chen

InstaDeep™

INTRODUCTION

Reinforcement learning in the multi-agent environment is difficult because of the changing policies of all the agents, which makes the environment not stationary. In this research project, we tackle the non-stationary problem of the multi-agent system by introducing an intent inference component to model other agents. The inferred intent can replace the changing policies of other agents in the decision-making process so that agents can learn the best response[Bowling and Veloso, 2004].

METHOD

In our method, each agent maintains a record of other agents' history action trajectories, then build an embedding of it. Later, we feed the embedding along with the current observation into the state and the state-action value estimations.

We construct the intent inference component F that takes the other agents' history action trajectories τ^i as input to infer the intents of other agents. We get it by passing the action trajectories through an RNN network with LSTM cells. It outputs the implicit hidden intent embedding gi and then we construct the input si + gi to feed it into the policy π^i and value estimator Vi. Meanwhile, an auxiliary supervised learning task is introduced to make the learning of intents more accurate by using the implicitly hidden intent embedding to predict the next timestep's trajectories $\tau^{i`}$. With the implicit intent inference, the learning can be considered as in an approximative stationary environment with approximative best response learning.

Table 1: Statistics of the OoredooTn dataset.

Q function:

$$Q^i(s, a) = Q^i(s, a^i, a^{-i}) \approx Q^i(s, a^i, g^i)$$

Value function:

$$\hat{V}_t^i(s) = \sum_{a^i} \pi^i(a^i|s, g^i) Q_t^i(s, a^i, g^i)$$

The update rule of the I3 Q-function is defined as:

$$Q_{t+1}^i(s, a^i, g^i) = (1 - \alpha_t) Q_t^i(s, a^i, g^i) + \alpha_t [r_t^i + \gamma] \hat{V}_t^i(s`)$$

The advantage function is:

$$A_t^i(s, a^i, g^i) = Q_t^i(s, a^i, g^i) - V^i{}_t(s^i, g^i)$$

The optimization objective:

$$L(\theta_i) = \sum_{a^i} \pi^i(a^i|s, g^i) A_t^i(s, a^i, g^i)$$

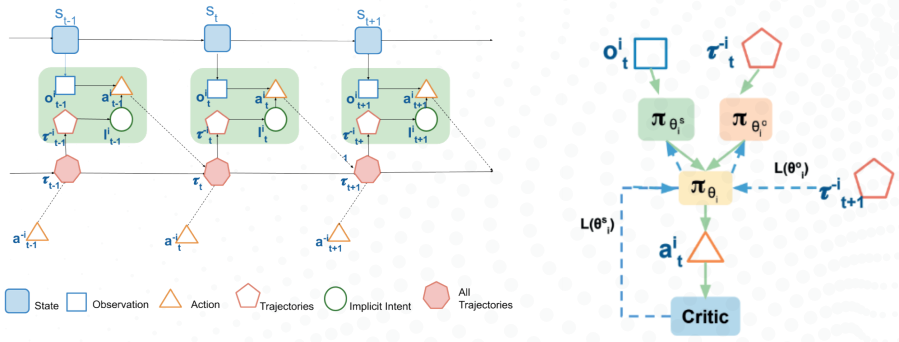

Figure 1: Architecture Diagram

In conclusion, with the implicit intent inference, the MARL problem is converted into a single agent best-response learning over the action distribution of other.

EXPERIMENTAL RESULTS

We tested I3 in two settings, the first one is stateless games, where simple RL methods like DDPG will fall into circular policy update deadlock. And we proved that I3 agents can achieve Nash Equilibrium. In Figure 2, The center point represents the Nash Equilibrium[Bowling and Veloso, 2004] solution of the game.

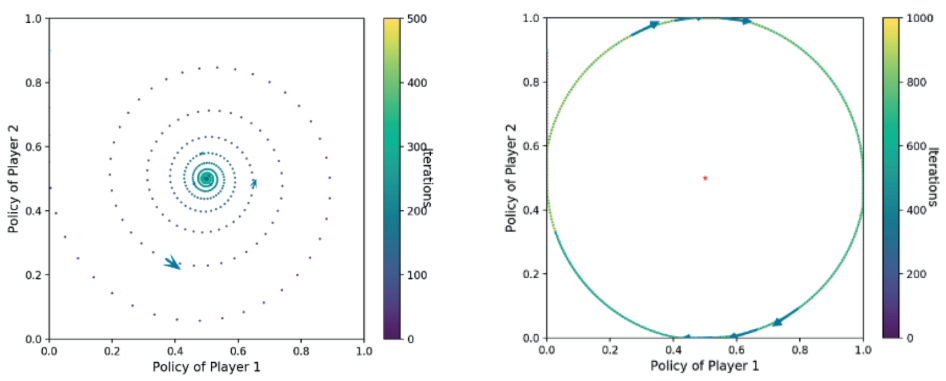

Figure 2: Policy Visualization for stateless two player game, Left I3, Right DDPG

The other experiments are in Multi-agent Particle Env.
Compare to the centralized critic method MADDPG[Lowe et al., 2017], we got competitive results on both cooperative and competitive games, and out-performed non-I3 version of the DDPG agent.

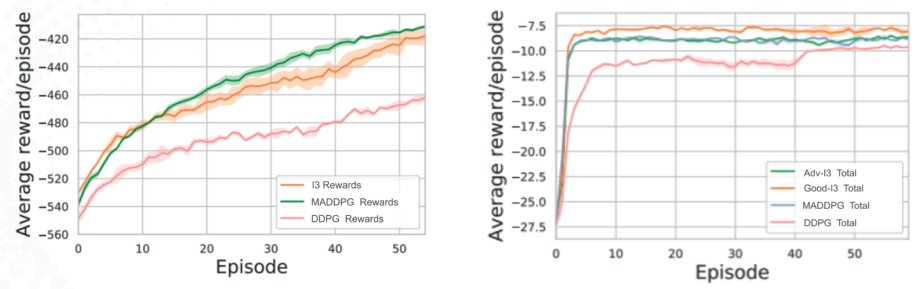

Figure 3: Multi-agent Particle Env experimental results.
Left: Spread(cooperative),
Right: Prey and Predator(competitive). In competitive setting, we tested I3 as prey and I3 as predator.

	Push Away (competitive)	Get to Landmark (Competitive)	Prey and Predator (Competitive)	Spread (cooperative)
I3 good	**-8.05 ±0.23**	4.19±1.6	**20.81±1.22**	**-410.81±3.255**
I3 adv	-8.63±0.23	**4.46±0.78**	16.22±2.22	-410.81±3.255
MADDPG	-8.84±0.16	3.58±0.82	19.59±1.78	-417.85-±13.69
DDPG	-9.64±0.15	3.09±1.7	6.34±0.87	-450.60±20.78

Conclusion

We presented the implicit intent inference multi-agent reinforcement learning methods, the non-centralized method enables agents to cooperate and compete with each other in a partially observable environment. The advantage of this method is that we don't need centralized training and it's fully distributed. This methodology can be applied to fields like robotic systems, and autonomous vehicles.

REFERENCES

[1] Daxiang Dong Yi Liu Ying Chen Wayne Xin Zhaoy Dianhai Yu Xiangyang Zhou, Lu Li and Hua Wu.
Multi-turn response selection for chatbots withh deep attention matching network. July 2018.
[2] Google Brain. Attention is all you need. Decembre 2017.
[3] Iulian V.Serban† Ryan Lowe, Nissan Pow and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. 2015.
[4] Chen Xing Zhoujun Li Ming Zhou Yu Wu, Wei Wu. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. 2017.