Emergent Social Learning via Multi agent Reinforcement Learning Summary

Objective of the paper:

To check if an independent RL agent can learn from other independent agents in a multi agent environment to improve their performance

Introduction and related work

Basic Model Free RL agents do not use social learning in most circumstances.
(Model Free RL : Agent cannot learn a function which models the environment).

By imposing certain conditions on the training environment and by using model based auxiliary loss ,generalized social policies can be obtained which enable agents to
1. Learn hard skills that cannot be learnt in single agent training
2. Adapt to new environment by taking input from agents in the new environment

In general, Model Free and imitation learning agents cannot perform properly when they are placed in new environment

Performance of an agent trained alone, then trained in multi agent environment, and deployed alone is greater than performance of agent that is trained alone and deployed alone

Humans and animals are good at social learning by the fact that they learn from others' experiences even without having direct access to them

In Imitation learning , it is needed to created explicit expert trajectories which are learnt by the agent where as in the Social learning , agent is able to find experts at tasks and learn from them

Two hypotheses that verified in the paper :
H1: Social learning helps an agent to learn complex policies that are difficult to learn when agent is trained alone

H2 : Social learning helps agent to adapt online

MARL (Mutli agent RL) environment is studied where the agents are independent of each other . They pursue their own rewards and do not get any reward for teaching or helping other agents
Autonomous driving is an example of such an environment
Vanilla model free RL agents cannot use social learning to improve their performance
in such environment

This paper proposes a training environment and model architecture that facilitates social learning

Interleaving training with experts and training alone , reliance on presence of experts can be reduced
This helps agents to perform well when deployed alone in the absence of experts

Agents trained in above mentioned manner perform well compared to agents trained alone and even than the agents trained using imitation learning

Imitiation learning teaches a particular skill
Social learning teaches agent to learn multiple skills from experts

Work in this paper is similar to third person imitation learning i.e. the agents are not given direct access to the expert's observations

Following is not assumed in this paper
A. Access to expert trajectories
B. Experts receive awards for teaching other agents

Following is considered in the paper:
A. Agents do not explicitly model other agents
B. Different from Inverse RL by the fact that reward function is not inferred . But learning from sub optimal experts is considered
C. Experts do not have privileged information about the environment anything that is not available to the novice agents

Paper analyzes why it is difficult for RL agents to learn from expert demonstrations in sparse reward environment and proposes to solve it

It uses model based auxiliary loss which helps to predict next state give current state to enable agents learn from trajectories where they did not receive award

Learning Social learning

Paper considers Multi-Agent Partially Observable Markov Decision Process (MA-POMDP) environments defined by the tuple <S, A, T, R, I, N>
S = state , A = Action , T = Transition Function R = Reward Function I= Inspection function ,
N = number of agents

$$J(\pi^k) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \, r_{t+1}^k \mid s_0\right],$$

discount factor $\gamma \in [0, 1]$. N

Each agent selfishly tries maximize the total expected discounted future reward

The agents which are trained independently, cannot directly observe other agents' observations or actions, and do not share parameters

An expert agent can demonstrate a novel state $s^{\sim}$ that is difficult to produce through random exploration. Ideally novice agents have to learn from this demonstration by updating their internal representation to model $s^{\sim}$

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) R_t$$

Along with demonstrated state , above is policy gradient objective used by the novice agent for learning the policy . Rt = total reward over the trajectory

$$R_t = \sum_{t'=t+1}^{T} \gamma^{t'-t-1} r_{t'}$$

In the episode where the demonstrated state occurred, if the novice fails to reach the goal , $r_{t'}$ = 0 . Then Rt' = 0
Policy objective is zero and no gradients are updated in the policy of the novice even though demonstrated state is given by the expert

To solve this problem ,temporal difference learning can be used temporarily

Consider Q-learning $Q(a,s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid a, s\right]$. As the agents continues to receive zero rewards , the Q(s,a) will become zero eventually

$$Q(\tilde{s}, a) = r + \gamma \max_{a'} Q(s', a') = 0 + \gamma 0 = 0$$

i.e for a novel state such as s˜ the Q value is zero

This makes learning from the cues of experts particularly difficult when novice does not reach the goal

Learning to model expert's policy will help novice to gain performance but it does not have access to the expert's state and actions .

Novice has to use it its own action and expert's policy , model the state transitions and there by gain knowledge of expert's policy . In the absence of external reward this will be difficult for novice

Social learning with auxiliary losses

Additional layers thetaA are added predict the next state given current state and they use unsupervised MAE auxiliary loss to train the network

$$\hat{s}_{t+1} = f_{\theta_A}(a_t, s_t); \qquad J = \frac{1}{T}\sum_{t=0}^{T} |s_{t+1} - \hat{s}_{t+1}|$$

Consider a demonstrated state is present in the trajectory and the auxiliary loss will not be zero unless the agent perfectly predict the novel demonstrated state . Non zero gradients can now be used by the novice agent to model the representation of the world

This architecture also implicitly improves the agent's ability to model other agents' policies, since it must correctly predict other agents actions to predict next state

This approach is called social learning with auxiliary loss

Modeling the other agent's policy does not force these independent agents to copy the actions of other agents (unlike in imitation learning)

PPO algorithm provided better performance with this approach and also GAE method is used to estimate the advantages

Agents use

- a convolution network to learn pixel representation from St
- As the environments that are non markov and partially observed are also investigated, LSTMs are used to model history of states or observations . LSTM hidden states are stored in the experience replay buffer

Social learning environments

Paper introduces a novel environment specifically designed to encourage social learning by making individual exploration difficult and expensive, and introducing prestige cues

In Goal Cycle , agents are rewarded for navigating between several goal tiles in a certain order, and penalized for deviating from that order. The goal tiles are placed randomly and are visually indistinguishable, so it is not possible for an agent to identify the correct traversal order without potentially incurring an exploration penalty .Exploration is made difficult in this manner

Without exploration penalty , agent will not be inclined to do social learning

Prestige cues are implemented in Goal Cycle signal through agents changing color as they collect rewards over the course of each episode

At time t, the color of an agent is a linear combination of the red and blue RGB color vectors: color$t$ = blue · c˜t +red· (1−c˜t), where c˜t = sigmoid(ct), and ct is a reward-dependent prestige value given by:

$$c_t = \begin{cases} \alpha_c c_{t-1} + r_t, & r_t \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Thus the color of each agent changes from red to blue as it collects rewards (slowly decaying with constant αc, but reverts to red if it incurs a penalty

Agents that perform well in high penalty environments using a curriculum: those are initially trained in low-penalty environments and gradually increase the penalty

Social learning in new environments:

Test the zero-shot transfer performance of agents pre-trained in 3-Goal Cycle in three new environments. The first is Goal Cycle with 4 goals.

Also , transfer is tested to 17x17 Four Rooms and 19x19 Maze environments which represent a challenging transfer tasks

Experiments

Compare the performance of agents with auxiliary predictive losses learning in an environment shared with experts (social ppo + aux pred) to that of agents with the same architecture but trained alone (Solo PPO + aux pred).

In 2nd experiment , an autoencoder is used and the reconstruction loss is calculated over it Social PPO + reconstruction loss is used and its performance is compared to (Social PPO + aux pred)

Also imitation learning performance is compared with social PPO + aux pred

Solo agents, even with the predictive auxiliary loss (solo ppo + aux pred), are not able to discover the strategy of reaching the goals in the correct order

Even when trained in the presence of experts, agents without an auxiliary loss (social ppo (vanilla)) failed in the same way

In contrast, social agents trained with auxiliary predictive or reconstructive losses were able to achieve higher performance. This confirms hypothesis H1,

During each episode, social learners wait and observe while the expert explores the goals to find the correct order. Only then do they follow the expert's cues.

Without social learning , 3-Goal experts transfer poorly to both 4-Goal Cycle and Maze.
In latter two environments, social learning agents achieve better performance on the transfer task than the original experts, suggesting social learning may be a generally beneficial strategy for improving transfer in RL

Taken together, these results provide support for hypothesis H2

The agent is able to retain good performance in solo 3-goal environments as well as 4-goal environments with experts, indicating that it is learning to opportunistically take advantage of expert cues while building individual expertise in the 3-goal environment. In fact, agents trained with social ppo + aux pred perform better in solo transfer environments than agents exclusively trained in the solo environment

Limitations and Future Work

Current work focuses on exploratory navigation tasks only .It has to extend to manipulative tasks

Summary conclusion:

Reading this research paper helped to understand the basic overview on how the policies of the novice agents are trained using the auxiliary loss of the expert agent.
Needed to get detailed understanding by checking the code especially on the goal tiles and goal cycles part i.e. how the penalties are calculated based on them

References

Link: Emergent Social Learning via Multi-agent Reinforcement Learning
https://arxiv.org/pdf/2010.00581.pdf