Theoretical Research in Planning Simulations using Reinforcement Learning

Fall 2021 Discovery Final Presentation

Jenny (Peiru) Xu

Berkeley
UNIVERSITY OF CALIFORNIA

Reinforcement Learning

- Machine learning training method based on **rewarding desired behaviors**
- **Agents** interact with the environment by **taking actions**
- Goal: **maximize the cumulative reward**

- **Reward system**, no supervisors, delayed feedback, **optimal control**
- Real-life examples: chess games (Alpha Go), Atari games, flying helicopters

Transfer Learning

- **Store knowledge** gained when solving one problem and apply to a different but related problem
- **Transfer information** from previously learned tasks for the learning of new tasks

- **Connection with reinforcement learning:** potential to significantly improve the sample efficiency of a RL agent

Transfer Learning in RL: Existing Methods

- **Policy distillation** by Google DeepMind
  - Extract the policy of a reinforcement learning agent and train a new network
  - Train the teacher model using DQN (deep Q-learning), generate pairs of observation states and teacher Q-values, and train student with supervised learning over the generated pairs

- Advantages: smaller network size, better and more efficient performance

Transfer Learning in RL: Existing Methods

- Dynamic **Automaton-Guided Reward Shaping** for Monte Carlo Tree Search
  - Represent objectives as automata in order to define novel reward shaping functions
  - Utilize automaton-guided reward shaping to facilitate transfer learning between different environments

- Advantages: deal with non-Markovian process, low dimensionality of the automaton, accelerate learning speed

Berkeley
UNIVERSITY OF CALIFORNIA

Automaton Distillation

- Motivation: two weaknesses of policy distillation
  - **Same state space** for teacher and student
  - **Markovian assumptions** of RL

- **Automaton distillation**
  - Transfer information from a teacher to a student by learning parameters over the transitions of the automaton in a surrogate environment and using those parameters to train a DQN in the target environment
  - Distill knowledge from a teacher automaton to a student DQN

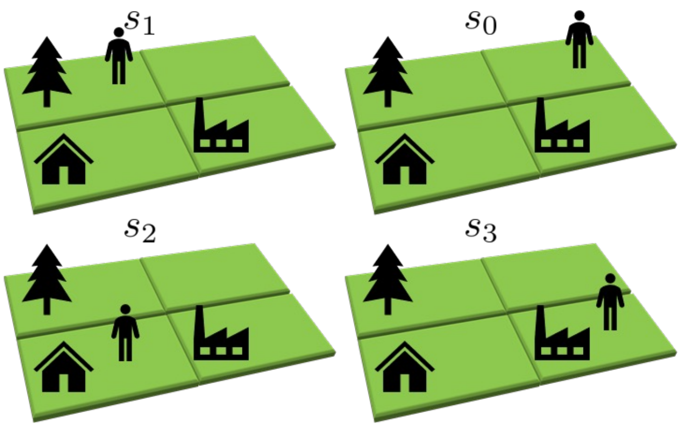
Preliminaries

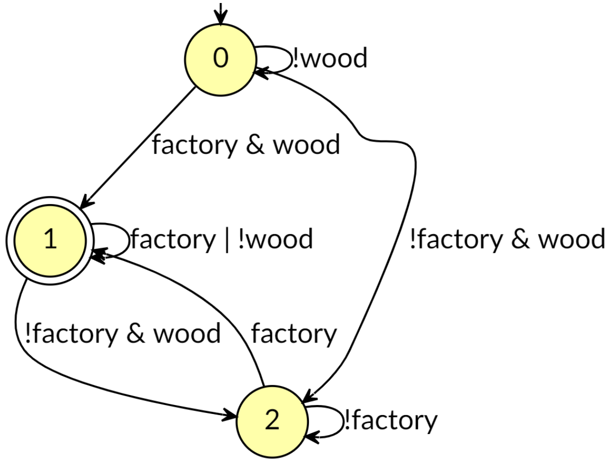
- **Non-Markovian Reward Decision Process**
  - Non-deterministic probabilistic process
  - $M = (S, s_0, A, T, R)$
  - S a set of states, s_0 initial state, A set of actions, $T(s'|s,a) \in [0,1]$ probability of transitioning, $R: S^* \to \mathbb{R}$ reward observed for a given trajectory of states, S^*: the set of possible state sequences

- Define a labeling function $\boldsymbol{L : S \to 2^{AP} = \Sigma}$
  - Map a state in the NMRDP to a set of **atomic propositions** in $\boldsymbol{AP}$ which hold for that given state.

Berkeley
UNIVERSITY OF CALIFORNIA

Illustration of NMRDP and Automaton





- Left: NMRDP consisting of four states, four actions
- Right: Automaton $A = (\Omega = \{\omega_0, \omega_1, \omega_2\}, \omega_0, \Sigma = 2^{\{wood, factory, house\}}, \delta = \left\{\omega_0 \xrightarrow{\neg wood} \omega_0, \ldots\right\}, F = \{\omega_1\})$
  - Objective: agent will eventually be on a tile containing wood and that, if the agent stands on said tile, then it must eventually reach a tile containing a factory.

Teacher DQN

- **Teacher DQN**
  - Standard reinforcement learning methods
  - Store samples of $((s, \omega), a, r, (s', \omega'))$ in the replay buffer ER
  - Define $\eta_{teacher}\colon \Omega \times \Sigma \to \mathbb{N}$:

$$\eta_{teacher}(\omega, \sigma) = |\{((s, \omega), a, r, (s', \omega')) \in ER | L(s') = \sigma\}|$$

  - Define $Q_{teacher}^{avg}\colon \Omega \times \Sigma \to \mathbb{R}$:

$$Q_{teacher}^{avg}(\omega, \sigma) = \frac{\sum_{\{((s,\omega),a,r,(s',\omega')) \in ER | L(s')=\sigma\}} Q_{teacher}(s, a)}{\eta_{teacher}(\omega, \sigma)}$$

Berkeley
UNIVERSITY OF CALIFORNIA

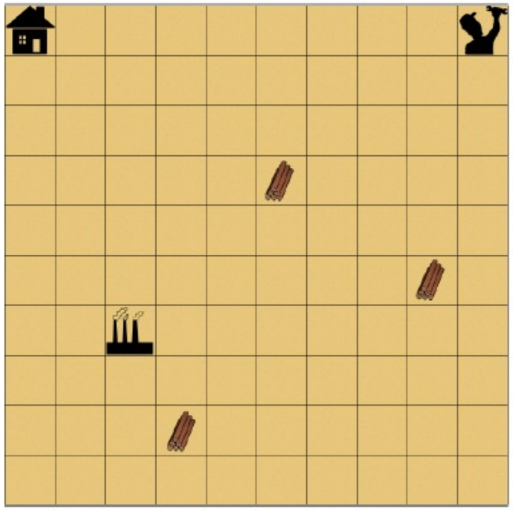
Student Loss Function

- **New loss function** for student training
  - Leverage the previous equations and the standard DQN loss function:
  - $Loss(\theta) = \mathbb{E}_{\left((s,\omega),a,r,(s',\omega')\right)\sim U(ER)}[\alpha\big(\omega, L(s')\big)Q^{avg}_{teacher}\big(\omega, L(s')\big) +$
    $\Big(1 - \alpha\big(\omega, L(s')\big)\Big)\Big(r + \gamma \max_{a'} Q(s', a'; \theta^{target}) - Q(s, a; \theta)\Big)^2]$
  - $\alpha: \Omega \times \Sigma \rightarrow [0,1]$: an annealing function
  - We use $\alpha(\omega, \sigma) = \rho^{\eta_{student}(\omega,\sigma)}$ where ρ=0.999

Berkeley
UNIVERSITY OF CALIFORNIA

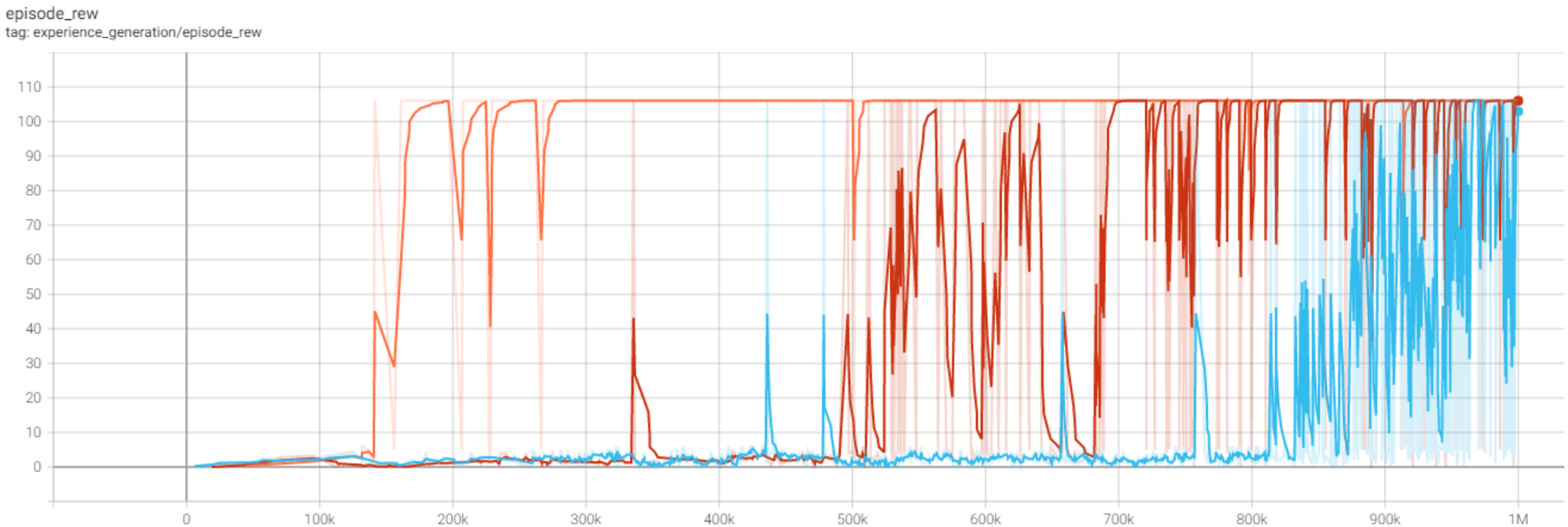
Experiment

- **Blind craftsman environment**
  - $AP = \{wood, home, factory, tools \geq 3\}$
  - Agent: first collect wood, bring it to the factory to make a tool from the wood. Hold a maximum of two woods at a time
  - Objective: craft at least three tools and arrive at home space
  - Linear temporal logic: $G(wood \Longrightarrow F\ factory) \wedge F(tools \geq 3 \wedge home)$



Results

- Compare three different RL algorithms
  - Automaton distillation from 7×7 to 10×10
  - Policy distillation from 10×10 to 10×10
  - DQN without transfer learning in 10×10



Berkeley
UNIVERSITY OF CALIFORNIA

Conclusions

- Automaton distillation: effective method of performing non-Markovian knowledge transfer
  - Automaton objective: much lower-dimensional representation of the environment
  - Encode non-Markovian reward signal
  - Can be performed for different teacher and student environment

- Next step: run over more complicated Non-Markovian environment, stabilize the automaton learning
- Submit paper to one of the top conference in the field

Data Science Insights

- Scientific impact: effective method of performing non-Markovian knowledge transfer in reinforcement learning environment
- Societal impact: feasible method to transfer learned information to slow simulation environments in Air Force(Advanced Framework for Simulation, Integration and Modeling, AFSIM)

- Insights in the field of RL: distillation from automaton gives more possibilities of transfer learning besides the standard DQN transfer

Thank you for listening!

Special thanks to:
Dr. Alvaro Velasquez, Arlo Malmberg, Ds-Discovery Program