Transfer Learning In Reinforcement Learning with Application to Robotics

COMP 767 Final Course Project

Monica Patel (260728093)

April 20, 2017

McGill University

Transfer Learning and Why

- RL agent learns by interacting with environment and gathering data.
- In robotics this is a physical agent and pretty expensive one!
- It can be best to learn in simulation and then use the knowledge in real world to avoid damage.
- Another scenario can be - Same task in dynamic environment.
- Transfer learning is technique to speed up the vanilla RL techniques.

Methods of Transfer

- Generalization and Mapping. How well the learned knowledge be generalized so that it can be used in target task. How well the source task maps to target task.
- What can be transferred?
  - Lower level knowledge such as $< s, a, r, a' >$ instances, action-value function, policies or complete model.
  - Higher level knowledge like partial policies - options, or reward shaping

Course Project Focus

- Course project focuses on two aspect of transfer - Lower lever knowledge transfer, source and target task mapping.
- Maze Navigation task - Goal position changed, Wall position changed.

Policy Reuse - (Fernandez and Veloso 2006)

π-reuse $(\Pi_{past}, K, H, \psi, \upsilon)$.

Initialize $Q^{\Pi_{new}}(s,a) = 0, \forall s \in \mathcal{S}, a \in \mathcal{A}$
For $k = 0$ to $K - 1$
    Set the initial state, s, randomly.
    Set $\psi_1 \leftarrow \psi$
    for $h = 1$ to H
        With a probability of ψ_h, $a = \Pi_{past}(s)$
        With a probability of $1 - \psi_h$, $a = \epsilon$-greedy$(\Pi_{new}(s))$
        Receive the next state s', and reward, $r_{k,h}$
        Update $Q^{\Pi_{new}}(s,a)$, and therefore, Π_{new}:
$$Q^{\Pi_{new}}(s,a) \leftarrow (1-\alpha)Q(s,a)^{\Pi_{new}} + \alpha[r + \gamma \max_{a'} Q^{\Pi_{new}}(s',a')]$$
        Set $\psi_{h+1} \leftarrow \psi_h \upsilon$
        Set $s \leftarrow s'$
$W = \frac{1}{K}\sum_{k=0}^{K}\sum_{h=0}^{H} \gamma^h r_{k,h}$
Return W, $Q^{\Pi_{new}}(s,a)$ and Π_{new}

Choosing from many Policies

- When Maze is same but the Goal location changes - Similarity measure function among policies.
  - Agent's task is to maximize average expected reinforcement per episode: $W = 1/K \sum_{k=0}^{K} \sum_{h=0}^{H} \gamma^h r_{k,h}$
  - For the different learned policies π_i the gain W_i is gain obtained when applying the $\pi - \textit{reuse}$ exploration strategy with policy π_i to learn policy π
  - Therefore when we have library of policies, the policy chosen for reuse is one that maximizes this gain while learning the target policy.
  - In PRQ-Learning algorithm this is done using softmax selection equation on all available policies:

$$P(\Pi_j) = \frac{e^{\tau W_j}}{\sum_{p=0}^{n} e^{\tau W_p}}$$

- When Walls are changed in the maze - Using domain knowledge, like image similarity measure like MSE.

Similarity Winner using Domain Knowledge

Figure 1: Willow garage world- Gazebo

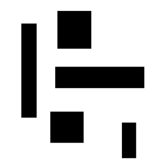

Figure 2: Tagert task

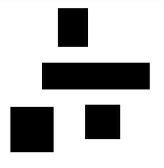

Figure 3: Source 1

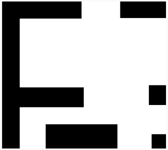

Figure 4: source 2

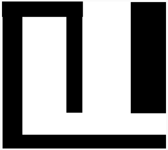

Figure 5: Source 4

Evaluation Metric and Results

Evaluation Metric and Results