Tell Agent Where to Go: Human Coaching for Accelerating Reinforcement Learning

Nakarin Suppakun  
Institute of Field roBotics, King Mongkut's University of Technology Thonburi  
Bangkok, 10140, Thailand

Suriya Natsupakpong  
Institute of Field roBotics, King Mongkut's University of Technology Thonburi  
Bangkok, 10140, Thailand

Thavida Maneeware  
Institute of Field roBotics, King Mongkut's University of Technology Thonburi  
Bangkok, 10140, Thailand
  E-mail: nakarin_sup@hotmail.com, suriya@fibo.kmutt.ac.th, prae@fibo.kmutt.ac.th
  www.kmutt.ac.th

Abstract

In this work, we proposed a method to accelerate learning by allowing a human to coach a robot behavior by inserting an intermediate target at the early phase of the reinforcement learning. By using an intermediate target, the different pair of policy and reward function was temporarily used to select an action that most likely to drive the robot toward the intermediate location, while the global reward function is still used for updating the state-action value. Q learning algorithm was used to test with the proposed method on three learning tasks: ball following, obstacle avoidance, and mountain car. The proposed technique resulted in better learning performance than the traditional RL.

Keywords: learning from demonstration, reinforcement learning, human assisted learning, semi-supervised learning, robot coaching

1. Introduction

In Reinforcement Learning (RL), long time training is usually required in order to train an agent to perform a given task. A large number of learning episodes is necessary to effectively propagate environmental rewards to each state action pair. However, instead of unnecessarily exploring states and actions freely, if there is a way that a human teacher could guide an agent toward the task completion at an early learning state, the agent could reach the goal state, and get the environmental reward within less learning iteration. In this work, a human coach guides an agent by placing a virtual intermediate target so that the agent would move toward. Another pair of policy and reward function would recognize this virtual target as a new goal state. A greedy policy is used in conjunction with an aggressive reward function until an agent reaches the virtual target. Meanwhile, the agent updates state-action value for every state-action encounter with the global environmental reward. After reaching the virtual target, the agent would be switched back to exploit and explore the states and actions through an e-greedy policy until next human
input is provided. In this work, a human only assists an agent in the early phase of learning process.

2. Related Work

Various techniques which using human inputs to assist an agent in learning a given task can be found in many literatures. In the work of A.L. Thomaz, and C. Breazeal1,2, a human teacher gives a scalar reward after a robot performs a task in a computer simulation. Human teacher was supposed to give the feedback after each action has taken; however, many people misunderstood and give the signal as if guiding what the robot should do next. In work of M. Hironaka and K. Suzuki3, a human coach would give a subjective evaluation as a binary feedback signal by modifying the reward function. There is also check for inconsistency of feedback and the signal is discarded if some inconsistency was found. There are also literatures about reward shaping technique4,5, the concept is that the feedback given by the human was used to modify the reward function by adding the interpreted value to the initial predefined reward (environmental reward). A list of certain feedback words was fixed and each word was predefined their corresponding value based on intensity level of meaning. In the work of S. Griffith et al.6, the policy shaping technique from an advice feedback was presented. However, they used a simulation feedback instead of an actual human feedback in order to control the consistency and likelihood of the feedback provided. W.B. Knox and P. Stone7-9 presented the system which allowed human teacher to give a binary feedback on the action taken. Human feedback model was trained by supervising technique and used for action selection as a policy. A variety of combinations on human input and traditional RL were tested8. The best combination was to use human input for an action selection only, which convince us to the same idea as well. In the work of W. D. Smart and L. P. Kaebbling10, they allowed a human to take over the policy for action selection on the early phase of learning, and then let the learning to proceed as usual later on. A human controlled the action directly via a joystick, while an agent updates its state-action value properly. In our work, a human coach assists the learning process by placing an intermediate virtual target to guide the agent toward the desired location. In this proposed method, there is no need for a human coach to manually select an action for each state-action that the agent encounters.

3. Methodology

In the proposed method, an agent performs both exploitation and exploration by $\epsilon$-greedy policy. (greedy with probability $1-\epsilon$ or random). Whenever a human coach assist an agent by giving an intermediate virtual target, a separated set of aggressive reward function is generated. With this reward function, a separated aggressive policy, which the action with the most reward received would be selected, is used. An agent would move toward the intermediate virtual target until the intermediate target is reached. In the meantime, each state-action encounter along the trajectory path would be updated the state-action value with a corresponding environmental reward (regardless of aggressive reward which was used solely for action selection). After reaching the guided target location, the agent would be switched back to $\epsilon$-greedy policy again until the next guide target is given by a human coach. The process is repeated until the global goal state is reached. In our experiment, the human guidance was only given in the early phase (the first 30 training episodes) and the traditional reinforcement learning was resumed later on.

4. Experimental Setup

All experiments were conducted in a computer simulation program. The first two tasks are in a robot soccer environment as shown in Fig. 1. The first task (ball following) is for the robot to move toward the ball. In the second task (obstacle avoidance), an obstacle (i.e. an opponent robot) was placed between the ball and the robot. From this program, a human coach could assign a virtual intermediate target for the robot by left clicking at the desired coordinate on the screen.

![Fig. 1. Simulation program for ball following and obstacle avoidance. Soccer field represents 6 x 4m.](image-url)

The third task is mountain car problem from Ref. 11. The purpose of this task is to drive a car back and forth in a valley between 2 mountains until the car can reach the
top of the mountain on the right. From this program, a human coach give a guided intermediate target for the car by left clicking at the screen. The horizontal position information would be used as a guidance input. Q reinforcement learning using a radial basis functions (RBFs) network as function approximator was used in our experiment (see Ref. 11). Information for each experimental task was explained as follow.

4.1. Ball following
State descriptors consist of the ball linear (Euclidean) distance, the sine and cosine of an angle from the robot heading to the ball, the sine and cosine of an angle to center of the opponent goal, and the linear distance from the opponent goal field line. List of actions are combination of linear and angular movement. Linear actions consist of go forward (0.1 m), stop, and go backward (-0.1 m). Angular actions consist of turn left (15 degrees), no turn (0 degree), and turn right (-15 degrees). Thus, we have the total of nine actions (3x3). Note that, the completely stop action (stop & no turn) is excluded.
A global reward function $r(s,a)$ was defined as follow. If the robot reaches the ball ( <= 0.15 m) and the ball is in front of the robot (in between angle -60 and 60 degrees), a positive reward +1.0 is given and the episode is successfully finished. If the ball is not in front of the robot, no score is given and the episode is failed. If the robot moves further away from the ball than the given threshold distance (6 m), a negative penalty score of -0.1 is given and the episode is failed. If the robot goes outside the field on either side, a penalty of -0.1 is given and the episode is failed. Otherwise, no score is given.
An aggressive reward function $f(s,a)$ extended the previous conditions by using the intermediate target given by a human coach instead of a global reward. If the current distance to the intermediate target is shorter than previous state’s, the positive score of +0.1 is added to the reward. If the current angular distance between the agent heading and the intermediate target is smaller than the previous state’s, the positive score of +0.1 is also added. The Greedy approach was used for the aggressive policy that is used in conjunction with the aggressive reward $f(s,a)$

4.2. Obstacle avoidance
All state descriptors explained in section 4.1 are also used with an addition of the obstacle linear (Euclidean) distance and the sine and cosine angle to the obstacle. The reward $r(s,a)$ condition was modified by including that if the robot collides with an obstacle (<= 0.3 m), then the penalty of -0.3 is given, and the episode is failed. List of actions, an aggressive reward function and policy are the same as in section 4.1.

4.3. Mountain car
State descriptors consist of the car position (in horizontal axis) and the velocity. List of actions consist of full throttle reverse, zero throttle, and full throttle forward. The reward function $r(s,a)$ was implemented differently from the original work11. A positive reward (+1.0) is only given when the car reaches the target location (top of the right mountain), and no reward for other cases. An aggressive reward function $f(s,a)$ was implemented separately as follow. If the current state is closer to the intermediate target than the previous state, a positive reward +0.1 is given. Again, greedy approach was used for an aggressive policy.

5. Results
Figs. 2, 3, and 4 illustrate the learning performance for the three tasks: ball following, obstacle avoidance, and mountain car, respectively. Each method and each task was repeated for 3 trials. In all three tasks, the guidance (coached) approach has higher accumulated rewards than the traditional RL.

![Fig. 2. A comparison learning performance for ball following.](image-url)
with reinforcement learning has accelerated the learning performance compared to the traditional RL.

References


6. Conclusion

The results showed that with the guided information from a human coach, an agent learn a task more effectively. An agent could use the guided information at the early learning episodes instead of exploring randomly at the beginning phase of the learning. As the agent reached the virtual intermediate target given by the human coach, the agent has performed what the human coach expected and also learned the list of useful actions along the way. The chance for exploration is still available as the policy is switched back to ϵ-greedy after the agent reached the intermediate target and possibly could find the even better solution. The experimental results from three different tasks also show that the proposed method that including an intermediate target from a human coach

Fig. 3. A comparison learning performance for obstacle avoidance.

Fig. 4. A comparison learning performance for mountain car.