Offline Reinforcement Learning-Based Human Guide Robot for Visually Impaired Navigation

Anushka Deshpande, Gil Gur-Arieh and Hadar Hai

Abstract

[sk: A trick I was taught: start by taking the first sentence of every paragraph of the introduction.]

Introduction

[sk:

- What is the problem you chose to investigate ?
- Why is it interesting ?
- Why is it hard ?
- What is a baseline approach to its solution / how has it been solved in the past
- What is your suggested approach ?
- What are the formal guarantees you provide How do you evaluate the approach ?
- What are the limitations of the suggested approach and how can it be extended ?

]

Example 1 *To illustrate a scenario in which an agent is assigned the task of guiding a visually impaired individual, we examine a situation where the agent (referred to as the guide robot) aims to guide the individual, connected to it by a leash, and navigate through a two-dimensional obstacle-filled map towards a goal, as depicted in Figure 1. In this scenario, the figure displays the agent represented in red and the visually impaired individual represented in blue. The agent possesses complete knowledge of the map layout, including obstacles and the location of the individual, throughout all time-steps. The agent must safely maneuver the individual through all obstacles to reach the goal while ensuring they do not collide with any obstacles. Conversely, the visually impaired individual lacks visibility of the map details and can solely perceive a directional signal aligned with the leash's direction, indicating the leash's maximum extension. Due to the human-like behavior and the lack of visual perception, the individual can sometimes act irrationally, influencing the situation unpredictably. Additionally, if the leash reaches its maximum length and the individual moves in the opposite direction, he can pull the agent, but the agent cannot tug the individual. This dynamic presents unique challenges for the agent in guiding the visually impaired individual effectively.*

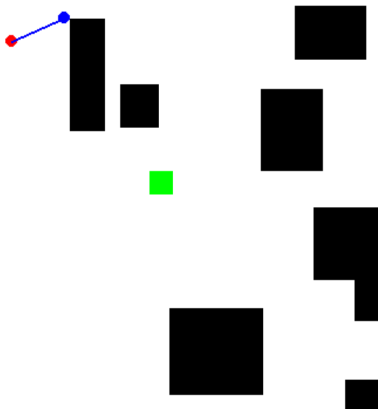

Figure 1: A simple 2D scenario featuring an agent (red) and a human (blue), connected by a leash, navigating through an obstacle-filled map towards a goal (green).

[sk: The contributions of this work are k-fold:

- new problem formulation
- new solution approach
- new formal guarantees
- performance improvement over existing appraoches
- new evaluation domains
- ...

]

Related Work

A framework with Offline Reinforcement Learning utilizing human-human interaction data was shown in (Hong, Levine, and Dragan 2023). In their work, Hong et. al. showed that by learning a dataset of suboptimal human-human interaction, an agent can learn to influence a human towards better performance, by modelling and learning the underlying human latent strategy. In this work, we focus on an agent assisting a blind human to achieve a common objective, where the focus is on the limited ability of the human to sense the world.

Background

[sk: The formal framework that is needed for understanding the problem statement and method]

Problem Statement

We explore a scenario where an agent (guide robot) is tasked with leading a visually impaired individual through a 2D obstacle map to a designated goal region, employing a leash, akin to the example depicted in Figure 1. We formulate our problem as a single-agent *Markov Decision Process* (MDP) which is defined by $< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R} >$, where $\mathcal{S}$ are the possible states, $\mathcal{A}$ represents the actions available to the agent, $\mathcal{T}\left(s'|s,a\right)$ denotes the transition function between states given an agent's action, and $\mathcal{R}\left(s,a,s'\right)$ signifies the reward accumulated by the agent. We do not utilize any discount factor ($\gamma = 1$).

This *Reinforcement Learning* (RL) problem aims to find an agent policy $\pi\left(a|s\right)$ which maximizes the cumulative reward:

$$R\left(s\left(t_{end}\right)\right)+\sum_{t=0}^{t_{end}-1}\left(R\left(s_t,\pi(s_t),s_{t+1}\right)\right)$$

States

The environment is represented as a 2D grid map, incorporating the guide robot, the human, the goal, and multiple obstacles. The state $s \in S$ is represented by an occupancy matrix M, and let M_{ij} represent the value of the cell at row i and column j. Each element M_{ij} of the matrix can take on the following values:

- $M_{ij} = 0$ if the cell at row i and column j is unoccupied.
- $M_{ij} = 1$ if the cell at row i and column j is occupied by an obstacle.
- $M_{ij} = 2$ if the cell at row i and column j is occupied by the guide robot.
- $M_{ij} = 3$ if the cell at row i and column j is occupied by the human.
- $M_{ij} = 4$ if the cell at row i and column j is a goal.

Actions

The guide robot can move horizontally, vertically, or remain stationary. The action space $\mathcal{A}$ is defined as follows:

$$\mathcal{A}=\begin{cases} ''U'': & (0,1),\\ ''D'': & (0,-1),\\ ''L'': & (1,0),\\ ''R'': & (-1,0),\\ ''S'': & (0,0)\end{cases}$$

Where each line represents a different action the agent can take, representing ”up”, ”down”, ”left”, ”right”, ”stationary” accordingly.

State-Action Transition Function

In our problem formulation, we assume the human can move in a similar manner to the guiding robot (up, down, left, right, stationary) and independently from the robot. It is important to emphasize that we formulate the human's actions as a part of the stochastic environment, as they are not directly controlled by the robot actions. Therefore, given a state $s \in \mathcal{S}$ and the robot's chosen action $a \in \mathcal{A}$, the subsequent state $s' \in \mathcal{S}$ (new occupancy matrix) will contain the new human and robot position, where the relation between the new human position and the robot's action (a) is unknown (and cannot be modeled directly):

$$\mathbf{h}_{t+1}=\mathbf{h}_t+\begin{cases}(0,1)\\(0,-1)\\(1,0)\\(-1,0)\\(0,0)\end{cases}$$

Where $\mathbf{h}_t$ denotes the human position (x,y) in time-step t.

The new robot position is defined by the robot action:

$$\mathbf{r}_{t+1}=\mathbf{r}_t+\begin{cases}(0,1) & a=''U''\\(0,-1) & a=''D''\\(1,0) & a=''L''\\(-1,0) & a=''R''\\(0,0) & a=''S''\end{cases}$$

Where $\mathbf{r}_t$ denotes the robot position (x,y) in time-step t. However, the above relation is only true unless one of the following conditions holds:

- The robot tries to move outside of map boundaries. In that case, $\mathbf{r}_{t+1} = \mathbf{r}_t$.
- The robot tries to move into an obstacle. In that case, $\mathbf{r}_{t+1} = \mathbf{r}_t$.
- The leash is at its maximum length and the robot tries to move away from the human. In that case, $\mathbf{r}_{t+1} = \mathbf{r}_t$.
- The leash is at its maximum length and the human moves away from the robot (the human drags the robot away). In that case, $\mathbf{r}_{t+1}$ will be the closest location to $\mathbf{r}_t$ which satisfies the euclidean distance between the human and the robot is no greater than the maximum leash size. Mathematically,

  $$\mathbf{r}_{t+1}=\arg\min_{\mathbf{r}}\|\mathbf{r}-\mathbf{r}_t\|$$

  Subject to the condition:

  $$\|\mathbf{r}_{t+1}-\mathbf{h}_{t+1}\| \leq \text{max\_leash\_size}$$

Reward

The reward accumulated by the agent at each step is the sum of the following components:

- $R_{human\_col}(s)$ - Negative reward for human collision with obstacle
- $R_{robot\_col}(s)$ - Negative reward for robot collision with obstacle
- $R_{goal}(s)$ - Positive reward for reaching the goal region
- $R_{step}(s)$ - Negative fixed reward for each time-step elapsed

YOUR METHOD

[sk: don't forget that this is a solution to the problem you stated in the previous section]

Formal Analysis

[sk: what are the formal guarantee that can be provided for your method:

- **Completeness:** does the algorithm find a solution when there is one ?
- **Soundness:** is a solution is found, is it a valid solution ?
- **Optimality:** is the solution a minimal-cost solution ?
- **Time complexity:** how long does it take to find a solution (in terms of node expansions)?
- **Space complexity:** how much memory is needed to perform the search (in terms of nodes and edges preserved in memory)?

]

Empirical Evaluation

In this work, a robot agent was trained using Offline Reinforcement Learning, utilizing human-human interaction data, in order to guide a blind human through an obstacle map. Empirical factors to evaluate the "success" of an agent might be, for example, Rate of episodes with success reaching the goal: $\frac{n_{\text{success}}}{N}$.

Dataset

To achieve the study's objective, we propose a data collection approach where two human participants, one representing the guide robot and the other the visually impaired individual, interact within a simulated environment by an interactive game. In this setup, the guide robot will navigate through the environment, while the visually impaired human will rely on its guidance. The guide robot player possesses complete knowledge of the state, whereas the visually impaired individual is only informed when the leash reaches its maximum length. At that point, the individual receives an arrow indicating the direction of the pulling force exerted by the leash, and according to this information, the individual player chooses where to move.

The dataset used for training, denoted as $\mathcal{D}$, comprises human-human interaction data collected during the interactive navigation tasks. Each data sample in the dataset consists of the following components:

- **State** (s_t): Representing the state of the environment at time t, including the positions of the guide robot $(\mathbf{r}_t)$, the visually impaired human $(\mathbf{h}_t)$, obstacles $(\mathbf{o}_t)$, and the goal $(\mathbf{g}_t)$. Mathematically, $s_t = (\mathbf{r}_t, \mathbf{h}_t, \mathbf{o}_t, \mathbf{g}_t)$. This information is extracted from the occupancy map at time t M_t of the environment.
- **Action** (a_t): Indicating the action taken by the guide robot based on the observed state and the action taken by the visually impaired human based on their limited information over time. Mathematically, $a_t = (\mathbf{a}_{r,t}, \mathbf{a}_{h,t})$, where $\mathbf{a}_{r,t}$ represents the action taken by the guide robot and $\mathbf{a}_{h,t}$ represents the action taken by the visually impaired human.

The dataset captures a variety of scenarios encountered during navigation, including different environmental layouts, obstacle configurations, and navigation strategies. It serves as the foundation for training the reinforcement learning agent to effectively guide visually impaired individuals in diverse environments.

Setup

The resulting agents (policies) will be tested in live experiments, in a similar setting which was used for the human-human data collection, with the agent replacing the guide robot player and choosing its actions according to the learned policy.

The baseline to which the agent's performance will be compared is a "naive" agent which is unaware of the leash or the human, and chooses its actions so that it reaches the goal in a minimum number of time-steps.

In terms of computational resources, we anticipate employing standard CPUs or, if necessary, GPUs. We will adjust this section accordingly.

Results

[sk: describe the results you received - the tables and diagrams and the key findings]

Discussion [optional, but highly recommended]

[sk: discuss what are potential explanations for the results you received and potential]

Conclusion

[sk: don't forget to mention limitations of your work!]

References

Hong, J.; Levine, S.; and Dragan, A. 2023. Learning to Influence Human Behavior with Offline Reinforcement Learning. arXiv:2303.02265.