Learning from Interventions with Hierarchical Policies for Safe Learning

Jing Bi[1], Vikas Dhiman[2], Tianyou Xiao[1], Chenliang Xu[1]
[1]{jbi5@ur., txiao3@u., chenliang.xu@}rochester.edu, [2]vdhiman@ucsd.edu

Summary

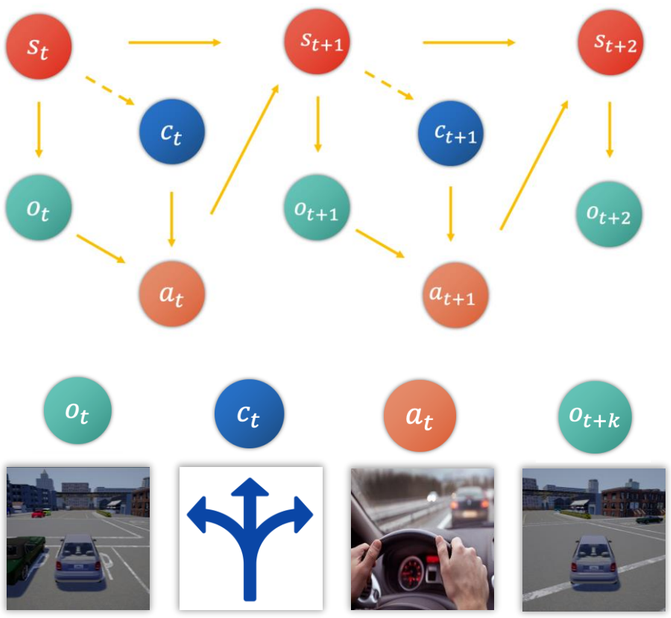
Motivation:

- Methods for solving compounding error often need to execute imperfect policy in the environment, which is infeasible in real world setting.
- The state-of-the-art Learning from Intervention fails to account for delay caused by the expert's reaction time and only learns short-term behavior.

Contributions:

- A new problem formulation of LfI that incorporates the expert's reaction delay.
- A novel algorithm combines LfI with Hierarchical policy
- A novel architecture to train the hierarchical policy
- An interpolation trick called Backtracking

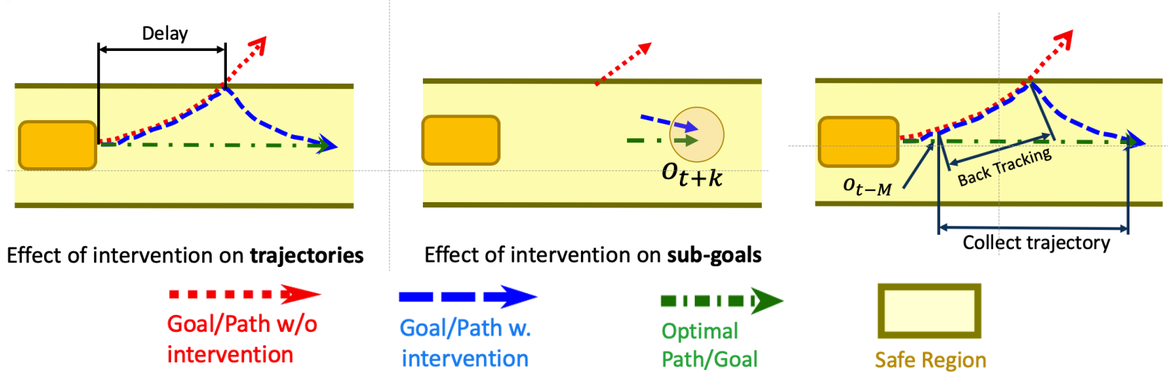
Problem Formulation



An agent interacting with an environment in discrete time steps which is considered as a Goal-conditioned MDP

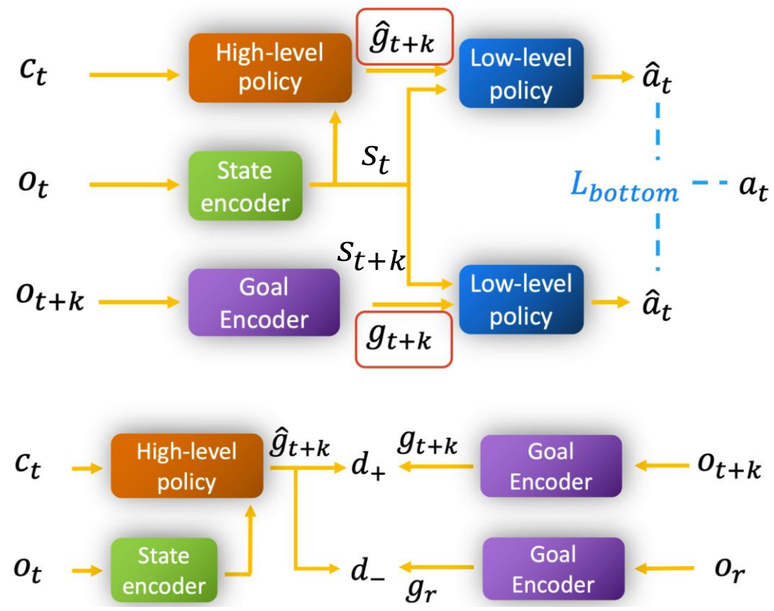
$\mathbf{o_t}$ = observation at time-step t
$\mathbf{a_t}$ = action based on o_t and c_t
$\mathbf{c_t}$ = command which indicates human intention
$\mathbf{o_{t+k}}$ = future observation after k time-steps

Methods



- Intervention with reaction caused a non-negligible difference between the intervened trajectory and the desired one.
- We formulate the intervention in terms of sub-goals to minimize this discrepancy ~~~ The top-level policy to predict the right goal and the bottom policy will generate the correct actions that will bring the car to the goal

Hierarchical Policies



The structure of hierarchical policy with triplet network

Result and Analysis

Figure 1: Top view of the map in CARLA simulator and real-world environment where experiments were conducted.

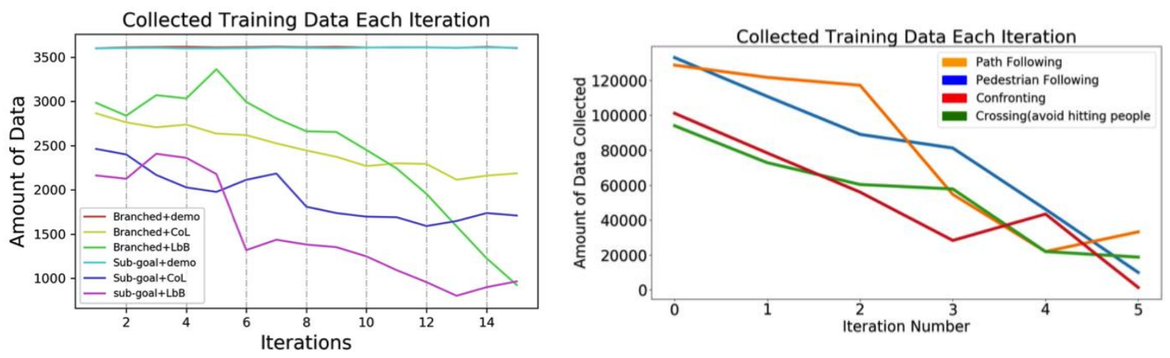


Figure 2: Comparison of the number of data-samples per iteration needed to train the various algorithms.

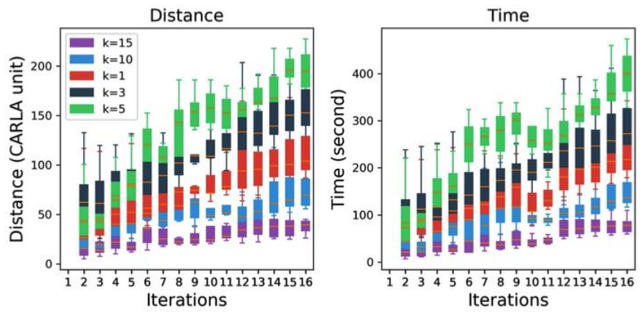


Figure 3: Evaluation of the effect of k on our proposed **Subgoal+LbB** algorithm.

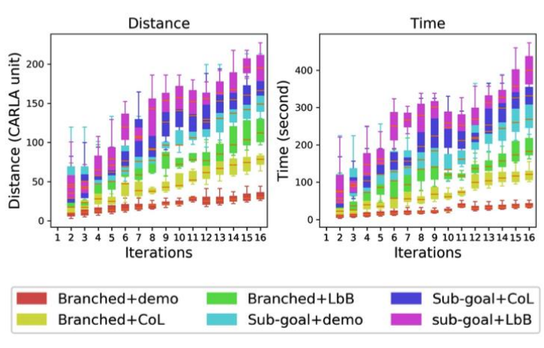


Figure 4: Distances and times without expert's intervention.

J. Bi, T. Xiao, and C. Xu are supported by NSF IIS 1741472, IIS 1813709, and NIST 60NANB17D191 (Sub). V. Dhiman is supported by the Army Research Laboratory - Distributed and Collaborative Intelligent Systems and Technology Collaborative Research Alliance (DCIST-CRA). This article solely reflects the opinions and conclusions of its authors but not the funding agents.

AAAI-20, New York