Human-Interactive Subgoal Supervision for Efficient Inverse Reinforcement Learning

Xinlei Pan
University of California, Berkeley
Berkeley, California, USA
xinleipan@berkeley.edu

Eshed Ohn-Bar
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
eshedohnbar@gmail.com

Nicholas Rhinehart
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
nrhineha@cs.cmu.edu

Yan Xu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
yxu2@andrew.cmu.edu

Yilin Shen
Samsung Research America
Mountain View, California, USA
yilin.shen@samsung.com

Kris M. Kitani
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
kkitani@cs.cmu.edu

AAMAS-2018

Maximum Entropy IRL

$$\theta^* = \arg\max_{\theta} \; -\sum_{d_i} p(d_i|\theta)\log\big(p(d_i|\theta)\big)$$
$$s.t. \quad f^{\pi} = \tilde{\mathrm{f}}^{D} \text{ (feature matching)}$$

其中 $\tilde{\mathrm{f}}^{D} = \frac{1}{N}\sum_{d_i \in D}\sum_{t=0}^{k} f_{it}$

$$f^{\pi} = \sum_{d_i} p(d_i|\theta) f_{d_i}$$

Notation

θ : parameter of reward function
π : optimal policy
d_i : trajectory
f_s : feature of state s
f_{d_i} : feature of trajectory
f^{π} : feature expectation of trajectory generated by policy
$\tilde{\mathrm{f}}^{D}$: feature expectation of trajectory given by human

IRL from Failure

$$\max_{\pi,w,z} H(D) + wz - \frac{\lambda}{2}\left||w|\right|^2$$

$$s.t.\quad f^{\pi} = \tilde{\mathrm{f}}^{D}$$

$$f^{\pi} - \tilde{\mathrm{f}}^{F} = z$$

Optimization step

$$\theta_d = \theta_d - \alpha\left(f^{\pi} - \tilde{\mathrm{f}}^{D}\right)$$

$$\theta_f = \frac{f^{\pi} - \tilde{\mathrm{f}}^{F}}{\lambda}$$

其中 $H(D)$ causal entropy is defined as:

$$H(D) = -\sum_{t}\sum_{s_{1:t},a_{1:t}} p(a_{1:t}, s_{1:t})\log(p(a_t|s_t))$$

Notation

$\tilde{\mathrm{f}}^{F}$: feature expectation of failure experience

z : difference between feature expectation of failure experience and feature expectation following π

w : lagrange multiplier of z

θ_d, θ_f: parameter of reward function learning from demonstration and failure experience respectively.

HI-IRL

- Interative setting
- Provide subgoal supervision-breaking task into subtasks.

Step1
Human expert provides several full demonstrations and define subgoals

Step2
Agents tries the defined subtasks. $s_r \rightarrow s_{sub_1} \rightarrow \cdots \rightarrow s_{sub_i} \rightarrow s_{goal}$.

Step3
Human provides further demonstrations if needed.

Step4
Learning reward functions from both failure experiences and expert demonstrations

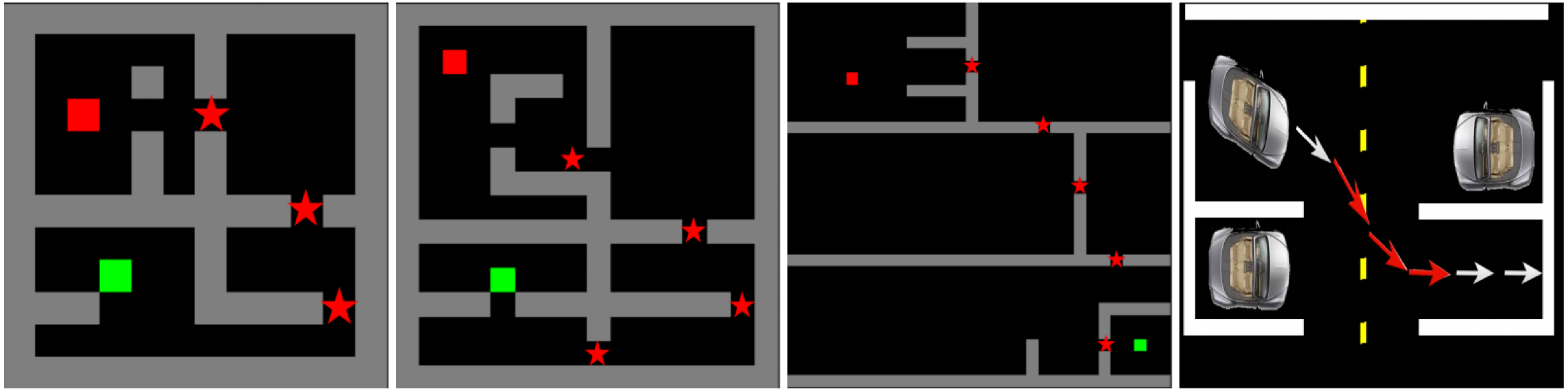
Algorithm

Algorithm 2 Human-Interactive Inverse Reinforcement Learning (HI-IRL)

Require: Set of initial demonstrations d_0, T, State Transition Matrix $\mathcal{T}$, θ^0, all state raw feature f, and human $\mathcal{H}$.
Return: Reward function θ^{T+1}
Define: $\mathcal{D}$: positive demonstrations; $\mathcal{F}$: failure experience; $\mathcal{E}$: agent experience; $\mathcal{S}_{sub}$: set of subgoal states
Start:

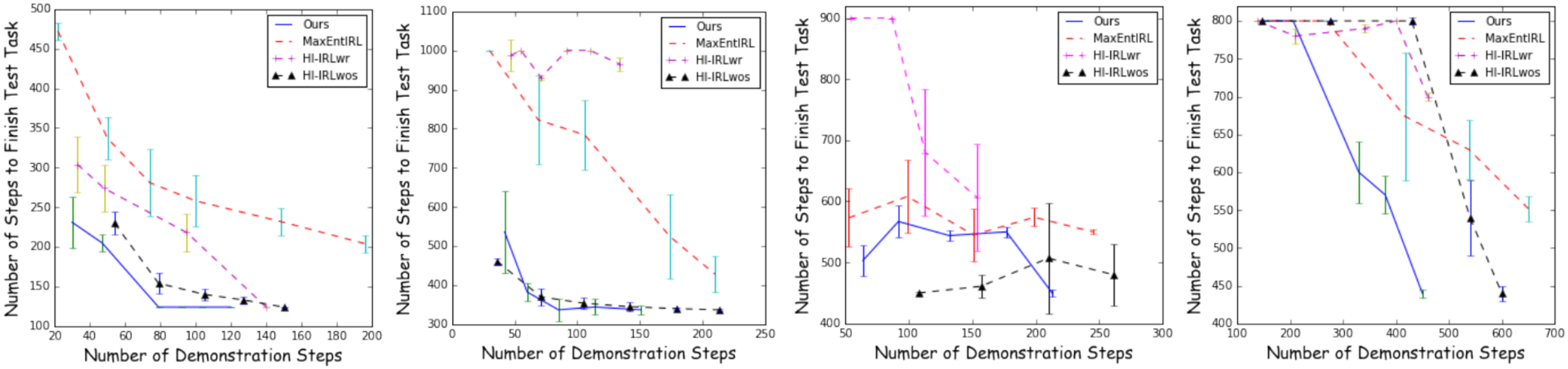
$\mathcal{S}_{sub}$ = specify_subgoals($\mathcal{H}$)
$\mathcal{D} = d_0$;
θ^1 = MaxEntIRL($\mathcal{D}$, θ^0)
for t $\in 1, 2, \ldots, T$
  $\mathcal{E} = \textsc{Rollout}(\theta^t, \mathcal{S}_{sub})$
  for e in $\mathcal{E}$
    $\mathcal{F}, \mathcal{D} = \textsc{UpdateDemo}(e, \theta^t, \mathcal{D})$
  $\theta_d^{t+1}, \theta_f^{t+1}$ = IRLFF($\mathcal{F}, \mathcal{D}, \mathcal{T}, \theta_d^t, f$) *(Alg. 1)*
  $\theta^{t+1} = (\theta_d^{t+1}, \theta_f^{t+1})$

Experiment



Baseline
1.MaxEntIRL
2.HI-IRLwos. The agent will be required to complete entire task and expert provide full demonstration.
3.HI-IRLwr. Randomly select states as subgoals

Result


Ours
MaxEntIRL
HI-IRLwr
HI-IRLwos
Number of Steps to Finish Test Task
Number of Demonstration Steps

(a) (b) (c) (d)