Bridging the Imitation Gap by Adaptive Insubordination

Luca Weihs*, Unnat Jain*, Iou-Jen Liu, Jordi Salvador,
Svetlana Lazebnik, Aniruddha Kembhavi, Alexander Schwing

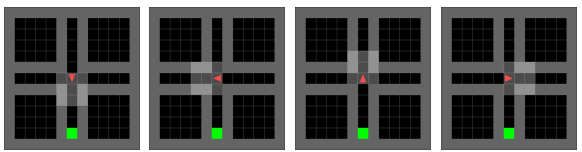
https://unnat.github.io/advisor

Privileged Experts and Mismatch With Students

Experts have access to full state
Student sees partial (egocentric) observations

Intuitive Example

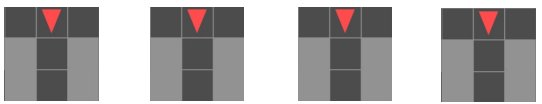
Expert sees:



Expert recommends:

Move forward	Turn left	Turn left/right	Turn right

But student's partial view makes these recommendations confusing...



Hence, student learns an averaged-policy:

Prop. 1: A student's policy is the average of the teacher's policy.

If $\pi^{\text{IL}} = \underset{\text{student policies } \pi}{\text{argmin}} \; \mathbb{E}[\, \text{CrossEntropy}(\pi, \pi^{\text{exp}}) \,]$

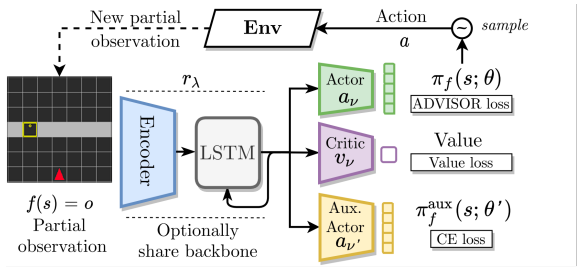
Then $\pi^{\text{IL}}(o) = \mathbb{E}[\, \pi^{\text{exp}}(S) \mid f(S) = o \,]$

Adaptive Insubordination (ADVISOR)

Idea:

1. Student estimates if it can imitate the expert
2. Accordingly, weigh IL & RL losses at each step

Schematic:

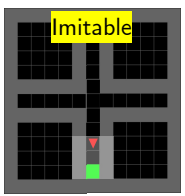


$$\mathcal{L}^{\text{ADV}}(\theta) = \mathbb{E}_\mu[w(S) \cdot CE(\pi^{\text{exp}}(S), \pi_f(S;\theta)) + (1 - w(S)) \cdot L(\theta, S)]$$

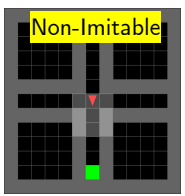
Auxiliary actor:

- Trained only by IL
- Estimates imitability of the current state

Outcome:



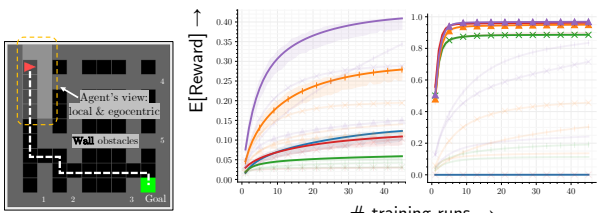
Small Aux. Actor Loss
High w(S) → More IL

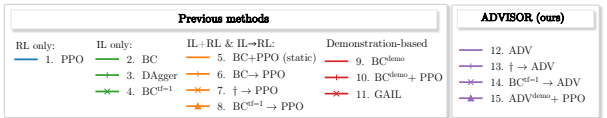


High Aux. Actor Loss
Low w(S) → More RL

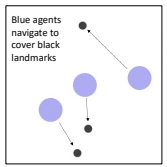
Results

Exhaustive evaluation in MiniGrid





Success across diverse tasks



CoopNav
(in MPE)

PointNav
(in AIHabitat)

ObjectNav
(in AI2-THOR)

Tasks → / Training routines ↓	PointGoal Navigation SPL @10%	PointGoal Navigation SPL @100%	ObjectGoal Navigation SPL @10%	ObjectGoal Navigation SPL @100%	Cooperative Navigation Reward @10%	Cooperative Navigation Reward @100%
RL only	30.9	54.7	6.7	13.1	−561.8	−456.0
IL only	30.1	68.7	3.8	9	−460.3	−416.7
IL + RL static	48.9	71.5	6.5	11.3	−475.5	−424.6
ADVISOR (ours)	**57.7**	**77.1**	**11.9**	**14.1**	**−419.9**	**−405.6**