Learning to Generalize from Sparse and Underspecified Rewards

Rishabh Agarwal, Chen Liang, Dale Schuurmans, Mohammad Norouzi

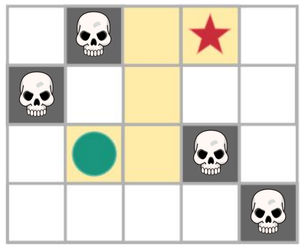
Problem

The success of RL algorithms hinges on the availability of **high-quality** and **dense** reward feedback.

How can we broaden the applicability of RL algorithms to real world environments with **sparse** and **underspecified** rewards?

Instruction Following

Language Instruction : "Right Up Up Right"



- **Correct program**

- **Spurious programs**

Given the input text instructions, the blindfolded agent (●) performs a sequence of actions, and only receives a reward of +1 if it reaches the goal (★).

Weakly Supervised Semantic Parsing

Question: Which nation won the most silver medals?

Rank	Nation	Gold	Silver	Bronze	Total
1	Nigeria	13	16	9	38
2	Kenya	12	10	7	29
3	Ethiopia	4	3	4	11
...	...	...	...	...	...
15	Madagascar	0	0	2	2
16	Tanzania	0	0	1	1
	Uganda	0	0	1	1

Converting natural language questions to SQL programs using only (question, answer) pairs.

- **Correct program**

(argmax *rows* "Silver")
(hop v_0 "Nation")

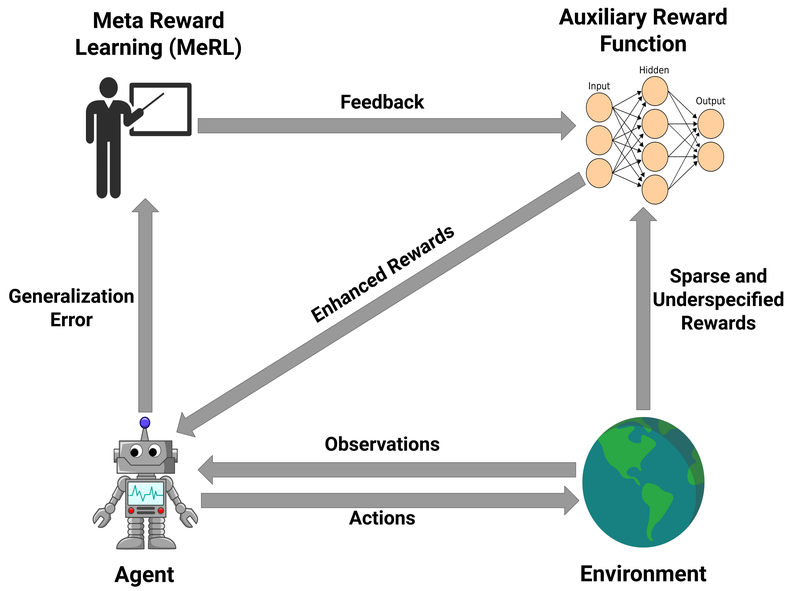
Right answer for the wrong reason

- **Spurious programs**

(argmax *rows* "Gold")
(hop v_0 "Nation")

(argmin *rows* "Rank")
(hop v_0 "Nation")

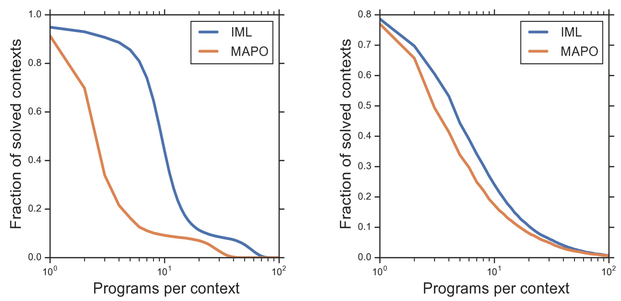
Meta Reward Learning (MeRL)



- Spurious trajectories are detrimental to generalization.
- Train the policy π_θ to maximize the auxiliary rewards R_ϕ
- Optimize auxiliary rewards to improve the generalization performance O_{val} of the policy π_θ

$$\theta' = \theta - \alpha \nabla_\theta O_{\text{train}}(\pi_\theta, R_\phi) \qquad \phi' = \phi - \nabla_\phi O_{\text{val}}(\pi_{\theta'})$$

Mode Covering Exploration

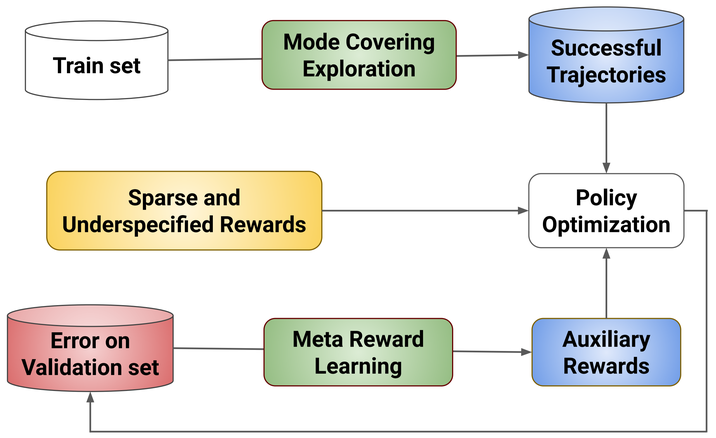


Fraction of contexts for which at least k programs are discovered during training using *IML* and *MAPO* on *WikiSQL* and *WikiTableQuestions* respectively.

- Disentangle exploration from exploitation
- IML is $D_{KL}(\pi^* \| \pi)$ while MAPO is $D_{KL}(\pi \| \pi^*)$
- Efficient exploration using IML (*mode covering* KL)
- Robust policy optimization using MAPO (*mode seeking* KL)

Our Contributions

- **Mode-covering exploration** to collect a diverse set of successful trajectories in a replay buffer.
- **Meta reward learning (MeRL)** to learn auxiliary rewards to disregard spurious trajectories.



Results

- *MAPOX* uses mode covering exploration on top of MAPO.
- *BoRL* is our Bayesian optimization approach for learning rewards.

Method	WikiSQL	WikiTable
MAPO	72.4 (± 0.3)	42.9 (± 0.5)
MAPOX	74.2 (± 0.4)	43.3 (± 0.4)
BoRL	74.2 (± 0.2)	43.8 (± 0.2)
MeRL	**74.8** (± 0.2)	**44.1** (± 0.2)

State-of-the-art results on *WikiTableQuestions* and *WikiSQL*, improving the upon prior work (*MAPO*) by **1.2%** and **2.4%**.

MeRL	MAPO

Question: Who was the first oldest living president?

(argmin *rows* oldest_living_president-date)
(hop v_0 "President")

(first *rows*)
(hop v_0 "President")

Question: What tree is the most dense in India?

(filter_str_contain_any *rows* [India] Location)
(argmax v_0 "Density") (hop v_1 "Tree_Name")

(argmax *rows*"Density")
(hop v_0 "Tree_Name")

Examples of programs generated on test queries of the *WikiTableQuestions* dataset.