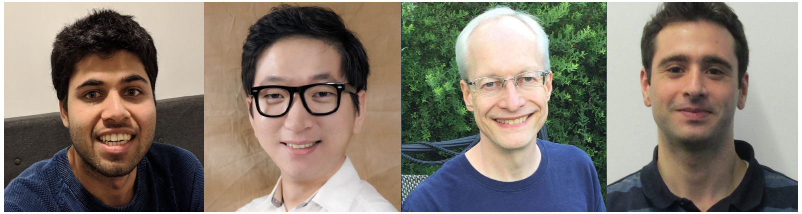
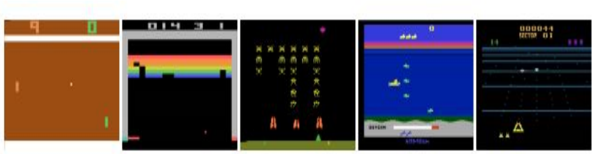
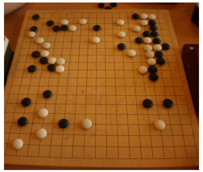
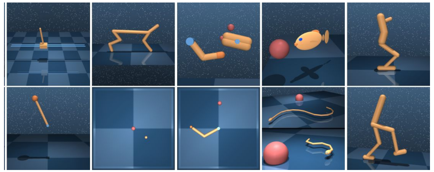
Learning to Generalize from Sparse and Underspecified Rewards

Rishabh Agarwal,
Chen Liang, Dale Schuurmans, Mohammad Norouzi



Motivation

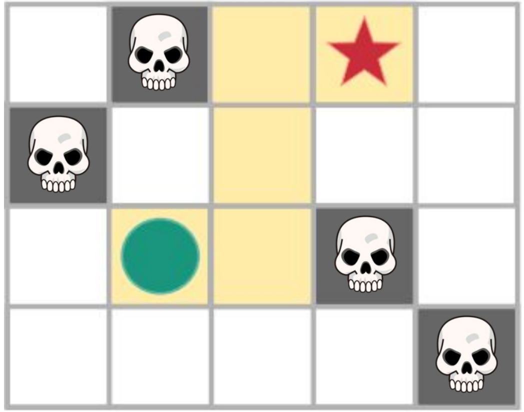
- ➢ Reinforcement learning has enabled remarkable advances:


- ➢ These advances hinge on the availability of high-quality and dense rewards.
- ➢ However, many real-world problems involve sparse and underspecified rewards.
- ➢ Language understanding tasks provide a natural way to investigate RL algorithms in such settings.

Instruction Following

Instruction: “Right Up Up Right”



: Blindfolded agent

: Goal

: Death

Possible Actions: ←, ↑, →, ↓

The reward is +1 if the goal is reached and 0 otherwise.

Weakly-supervised Semantic Parsing

Question: Which nation won the most number of 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

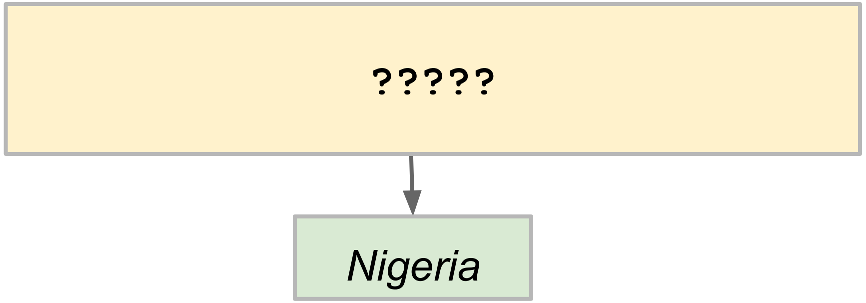
?????

Nigeria

Challenges: (1) Exploration, (2) Generalization

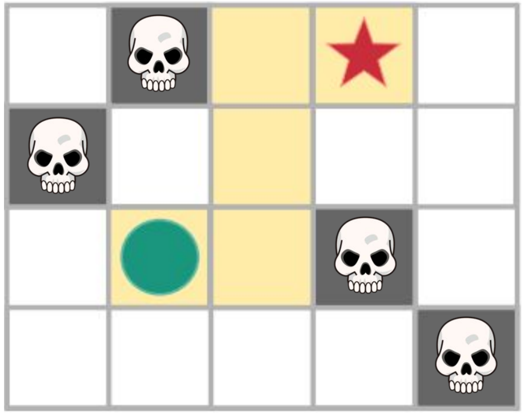
Question: Which nation won the most number of Silver medals?

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



Underspecified Rewards

Instruction: “Right Up Up Right”



Correct Action Sequence:

→ ↑ ↑ →

Spurious Action Sequences:

↑ → ↑ →

↑ → → ↑

↑ → ↑ ↓ → ↑

Underspecified Rewards

Question: Which nation won the most number of 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

```
v0 = (argmax all_rows r.Silver)
return (hop v0 r.Nation)
```

Nigeria

Underspecified Rewards

Question: Which nation won the most number of 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

```
v0 = (argmax all_rows r.Gold)
return (hop v0 r.Nation)
```

Nigeria

Underspecified Rewards

Question: Which nation won the most number of 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

```
v0 = (argmin all_rows r.Rank)
return (hop v0 r.Nation)
```

Nigeria

Underspecified Rewards

Recent interest in automated reward learning using expert demonstrations.

Learning Rewards without Demonstration

Recent interest in automated reward learning using expert demonstrations.

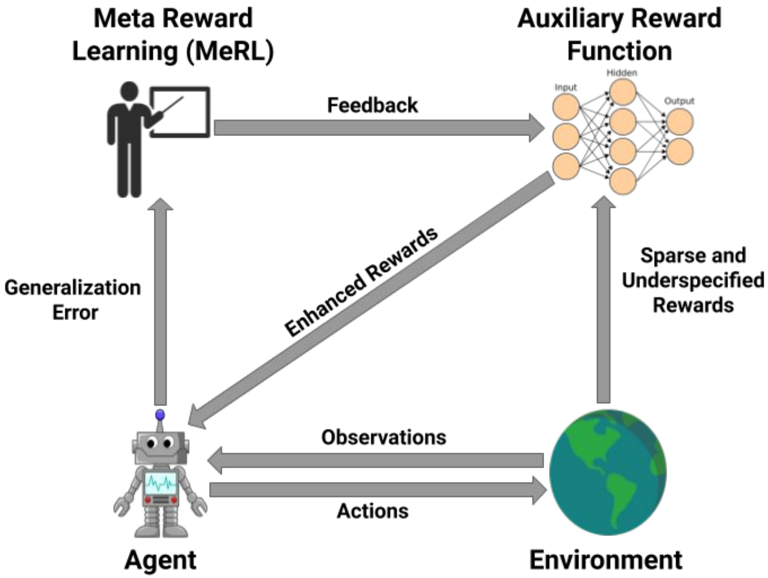
What if we don't have demonstrations?

Learning Rewards without Demonstration

Recent interest in automated reward learning using expert demonstrations.

Key idea: Use generalization error as the supervisory signal for learning rewards.

Meta Reward Learning (MeRL)



The auxiliary rewards R_ϕ are optimized based on the generalization performance O_{val} of a policy π_θ trained using the auxiliary rewards:

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

$$\phi' = \phi - \nabla_\phi O_{\text{val}}(\pi_{\theta'})$$

Tackling Sparse Rewards

- ➢ Disentangle exploration from exploitation.
- ➢ Mode covering direction of KL divergence to collect successful sequences .
- ➢ Mode seeking direction of KL divergence for robust optimization.

Results

➢ *MAPOX* uses *our mode covering* exploration strategy on top of prior work (*MAPO*).

Method	WikiSQL	WikiTable
MAPO	72.4 (± 0.3)	42.9 (± 0.5)
MAPOX	**74.2** (± 0.4)	**43.3** (± 0.4)

Results

- *MAPOX* uses *our mode covering* exploration strategy on top of prior work (*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)**

Results

- *MAPOX* uses *our mode covering* exploration strategy on top of prior work (*MAPO*).
- *BoRL* is our Bayesian optimization approach for learning rewards.
- *MeRL achieves* state-of-the-art results on *WikiTableQuestions* and *WikiSQL*, improving the upon *prior work* by **1.2%** and **2.4%**.

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)

Google

Poster **#49** tonight @Pacific Ballroom

bit.ly/merl2019