Task Agnostic Continual RL

Massimo Caccia, Jonas Mueller, Taesup Kim, Laurent Charlin, Rasool Fakoor





TL;DR:

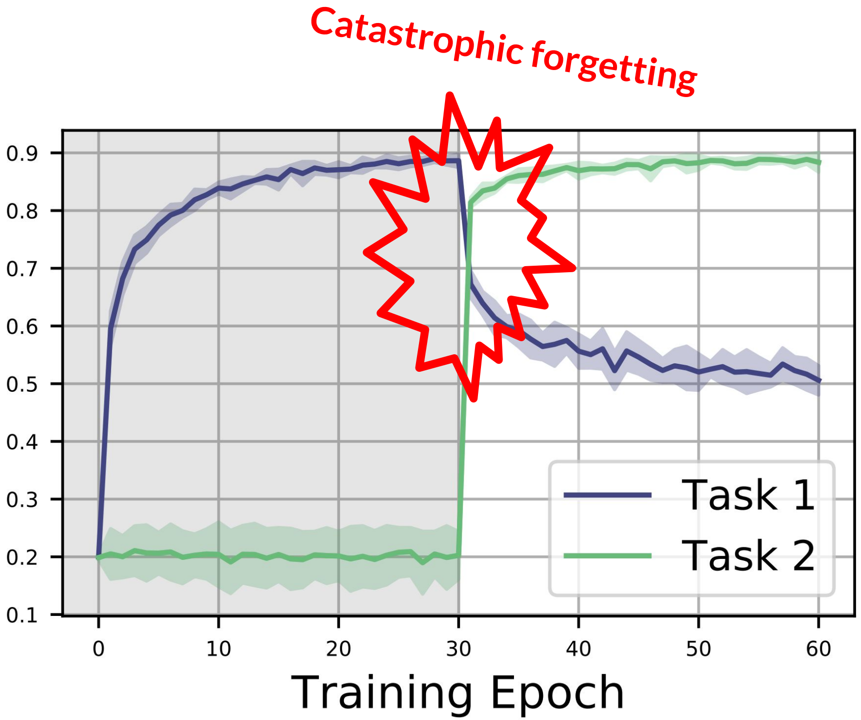
Overview

- Background
  - Continual Learning
  - Reinforcement Learning
- Task Agnostic Continual RL
  - Problem Statement
  - Soft upper Bounds: Task-awareness, Multi-task learning
- Methods
  - Backbone RL algo (SAC)
  - Baselines
- Empirical Findings
  - Benchmark
  - Task-agnostic > Task-Aware
  - Continual learning = multi-task learning
  - Hypothesis testing
- Discussion

Continual Learning (CL)

- Accumulating knowledge on non-stationary data distributions

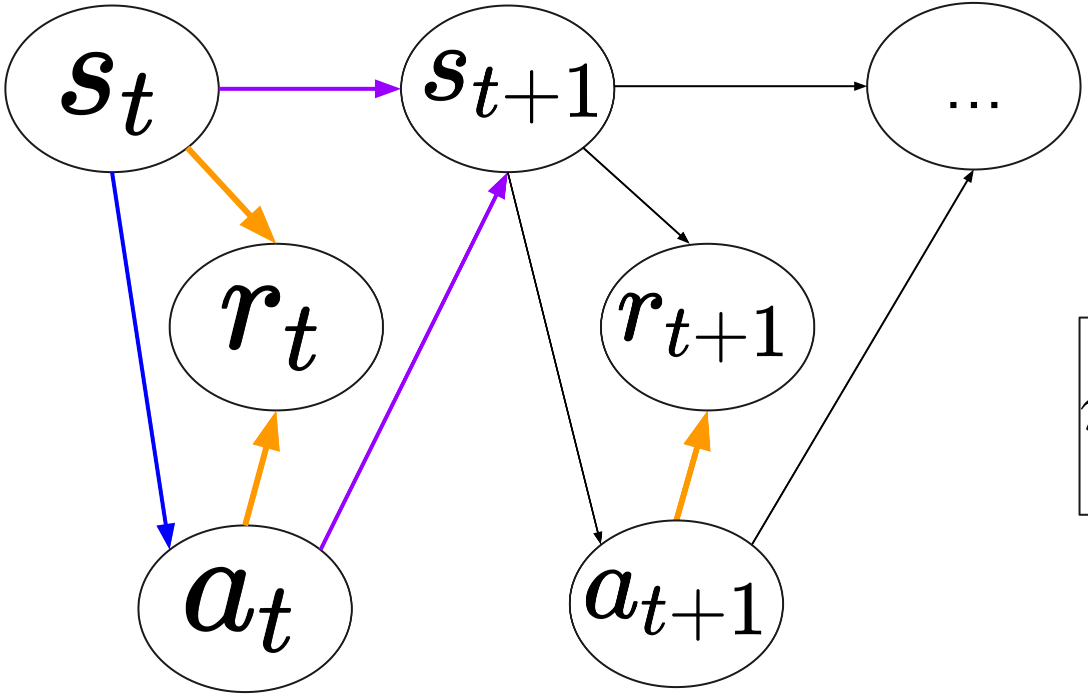
- ML/DL can't learn on changing data distributions (or tasks)



Continual Learning (CL): Why we care

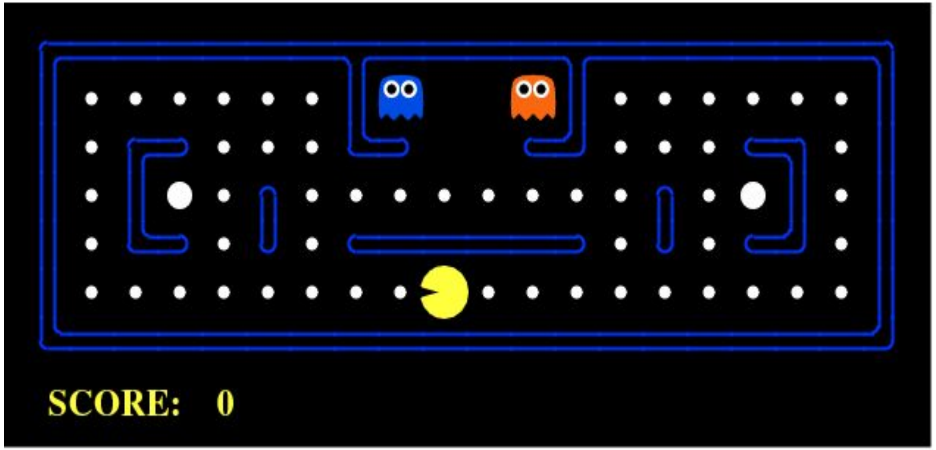
- industry/deployment argument
- Curriculum learning argument
  - Or continually increasing sample/compute efficiency
- Learning autonomously in an open world → AGI

Reinforcement Learning (RL) - MDP

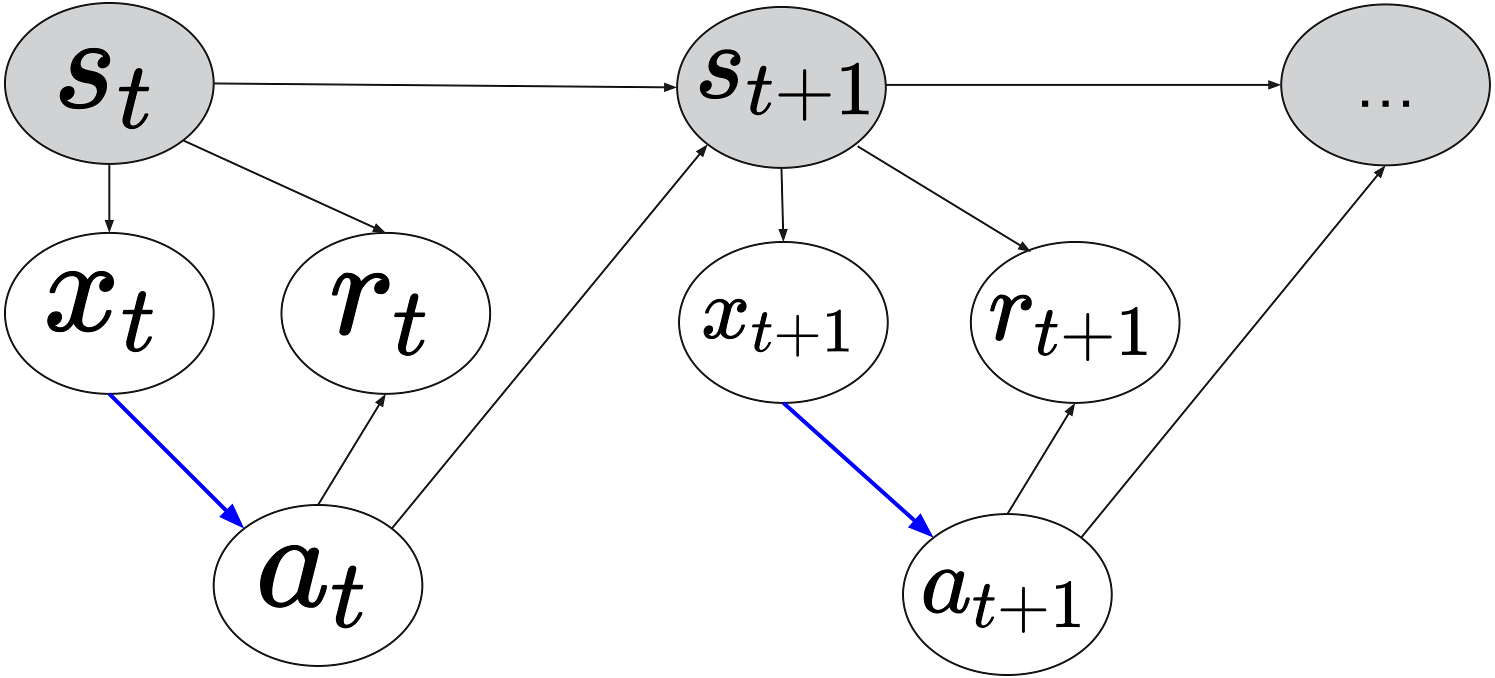


$$\boxed{\pi^* = \text{argmax}_\pi \mathbb{E} \sum_{t=0}^{\infty} \gamma^t r_t}$$

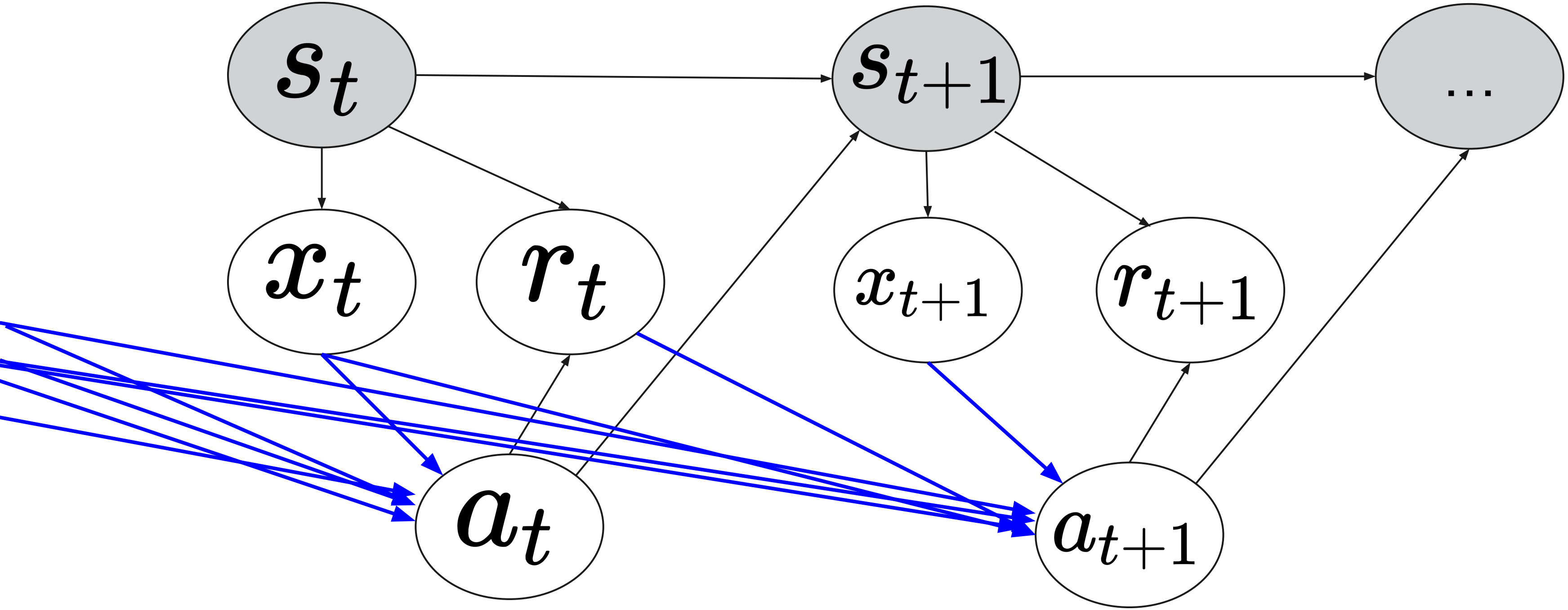
Reinforcement Learning (RL) - MDP



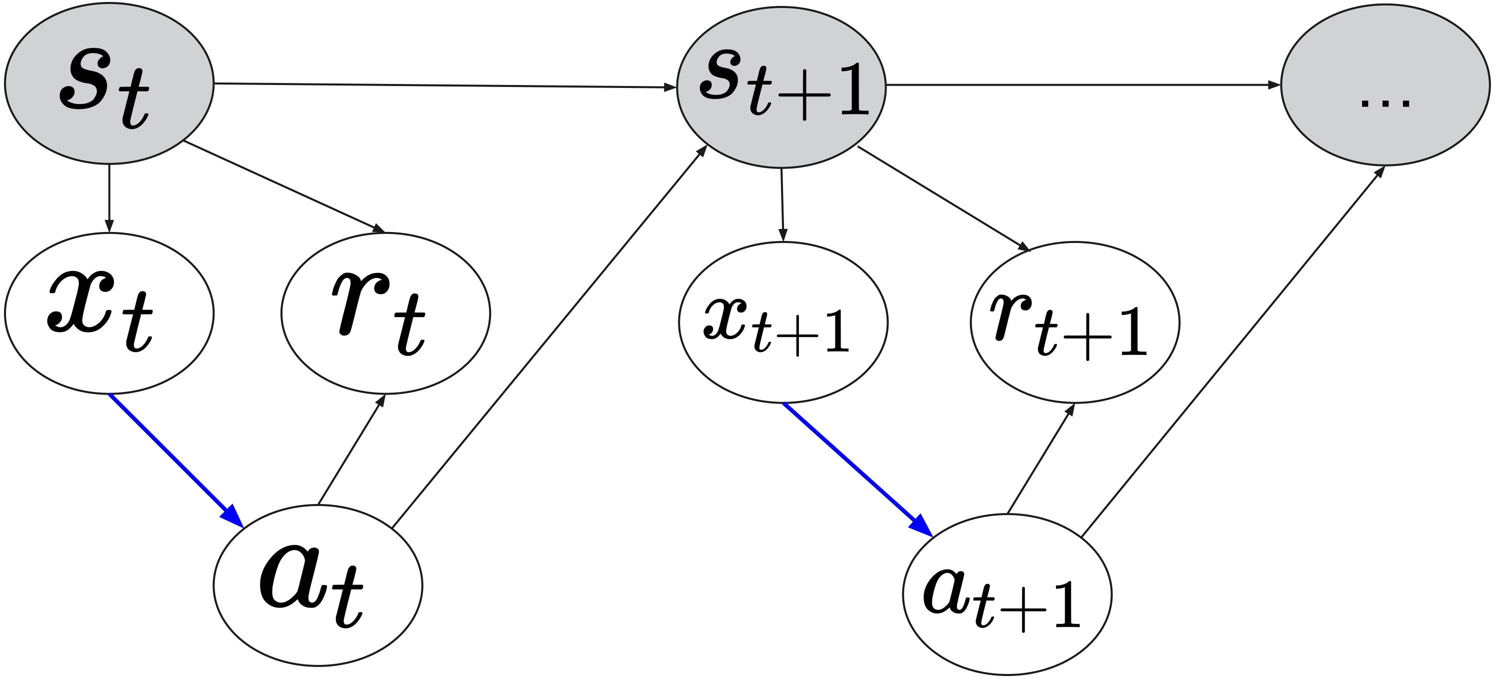
Reinforcement Learning (RL): POMDP



Reinforcement Learning (RL): POMDP

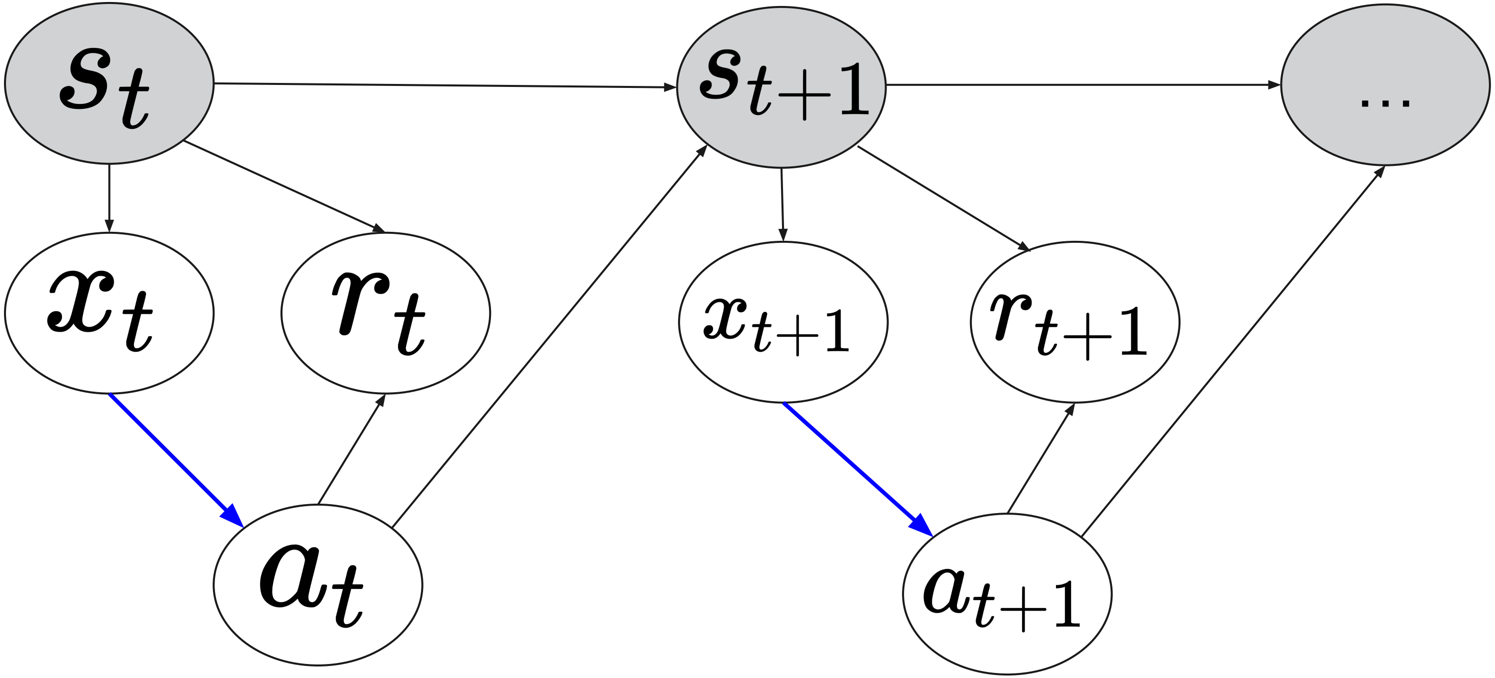


Reinforcement Learning (RL): POMDP



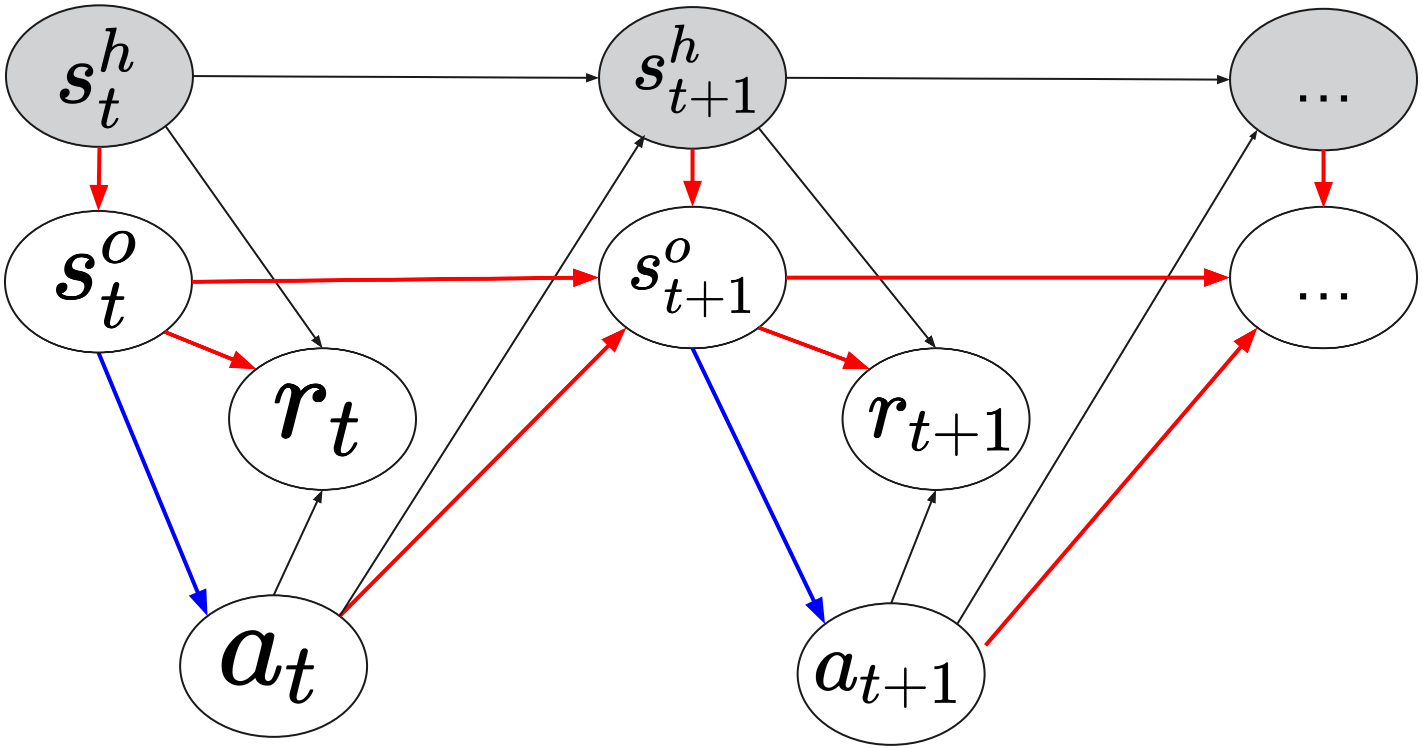
Reinforcement Learning (RL): - POMDP

TACRL as a POMDP special case

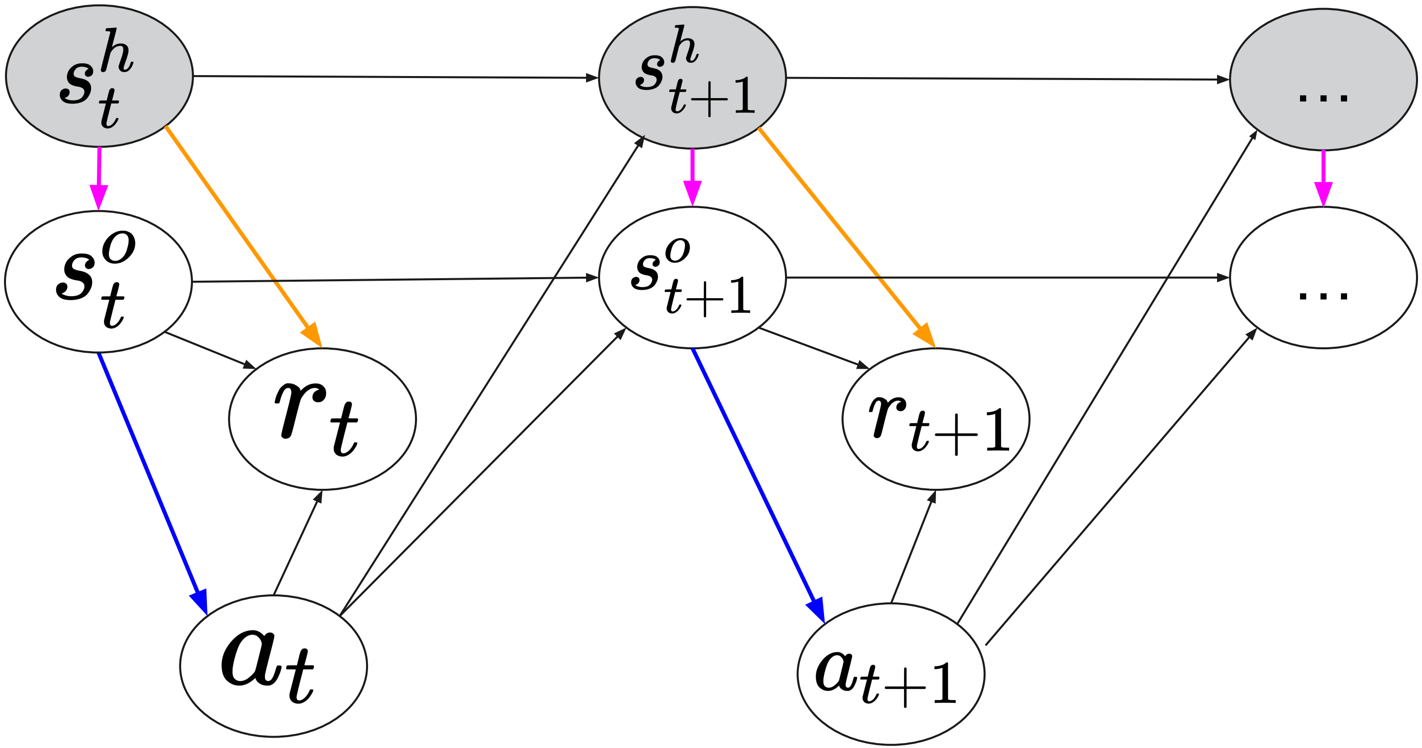


TACRL as a POMDP special case

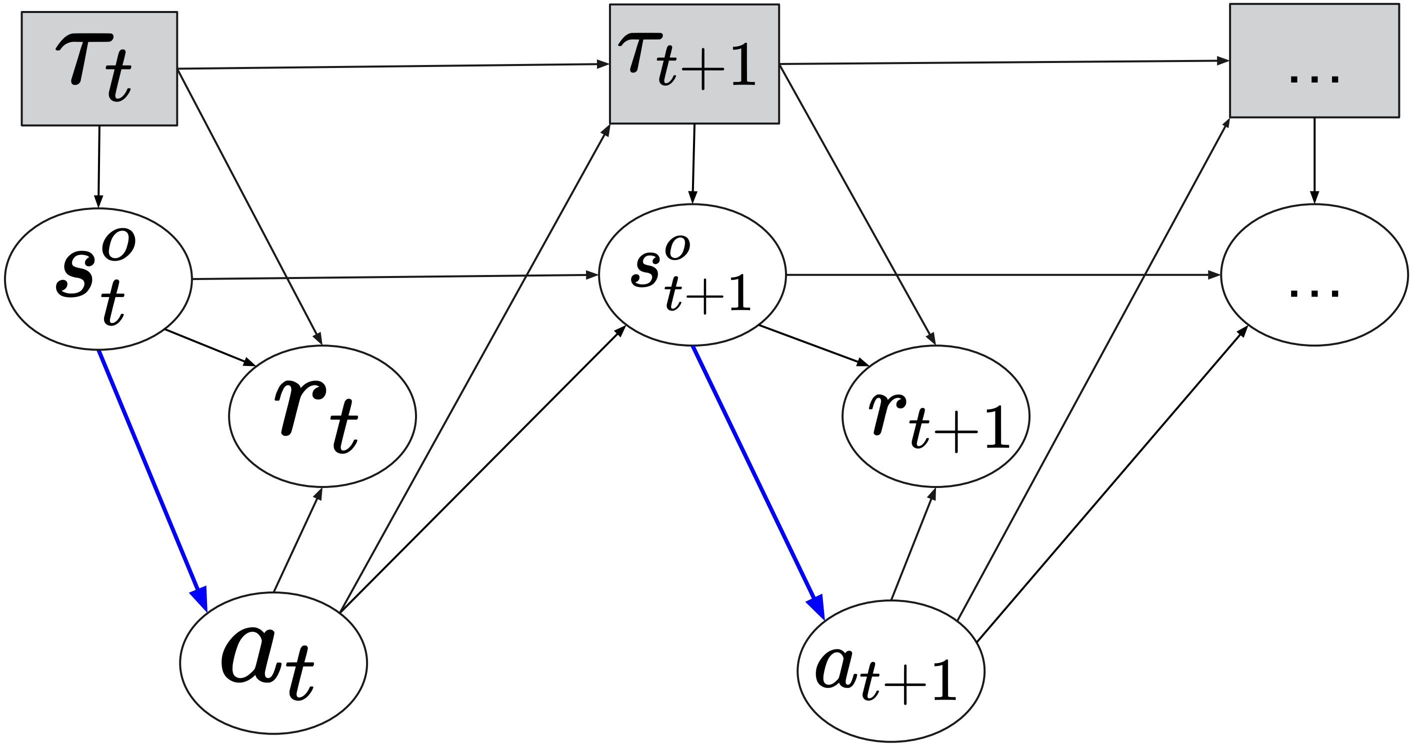
$$\boxed{p(s|a) = p(s^o|s^h, a)p(s^h|a)}$$



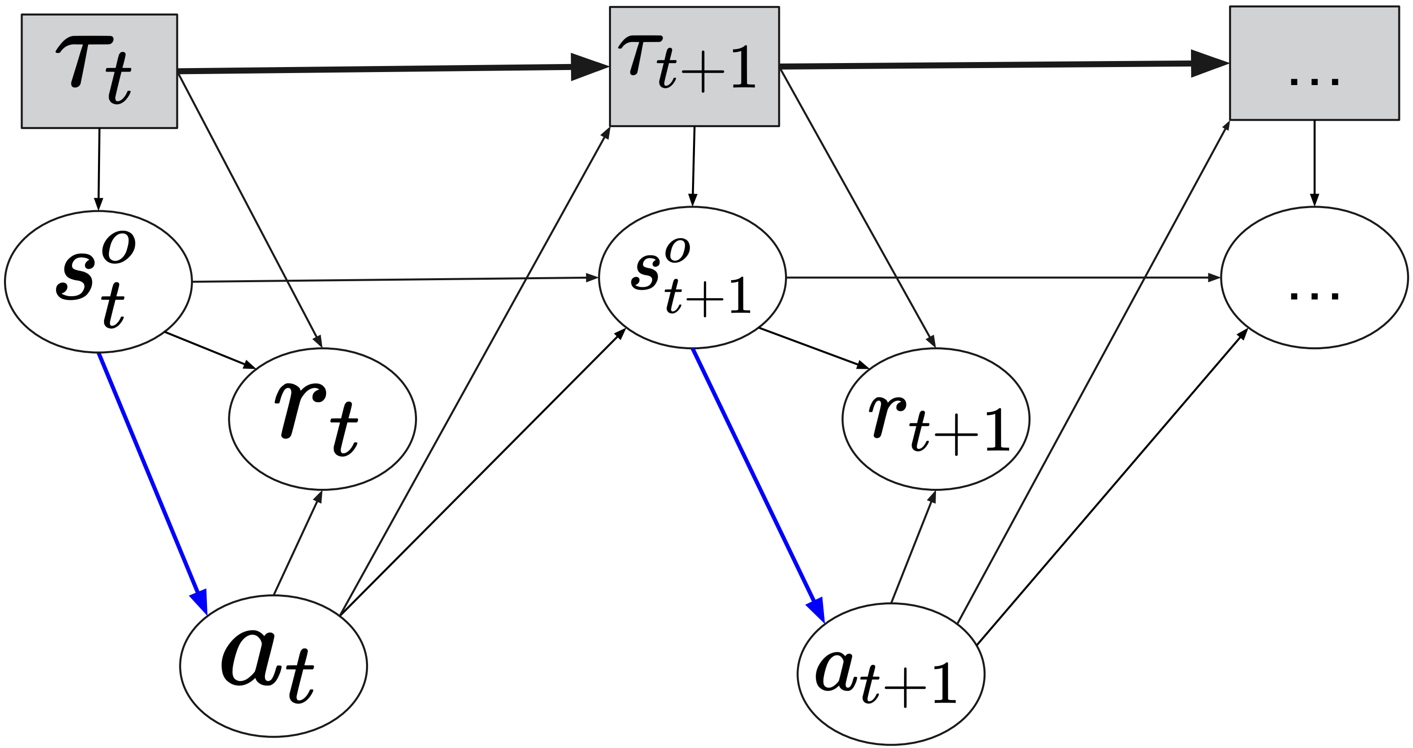
TACRL as a POMDP special case



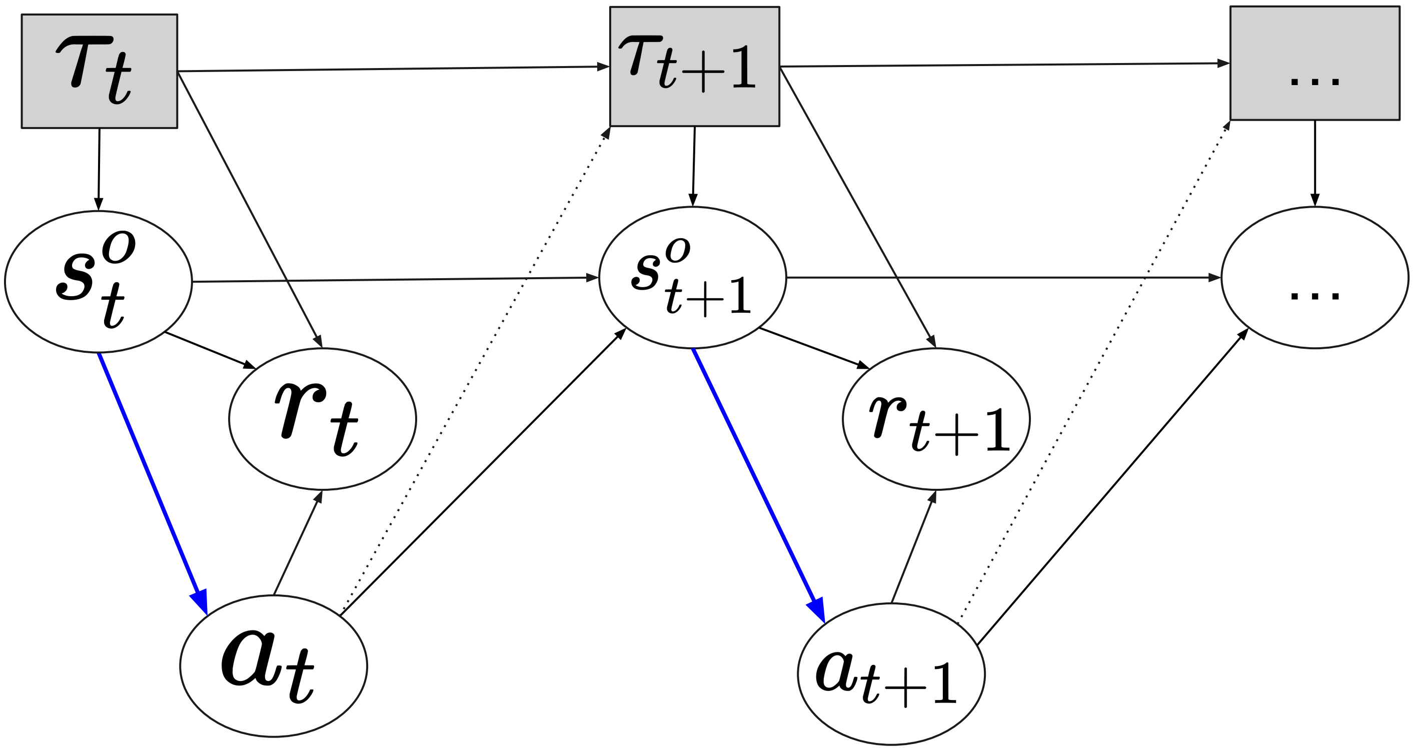
TACRL as a POMDP special case



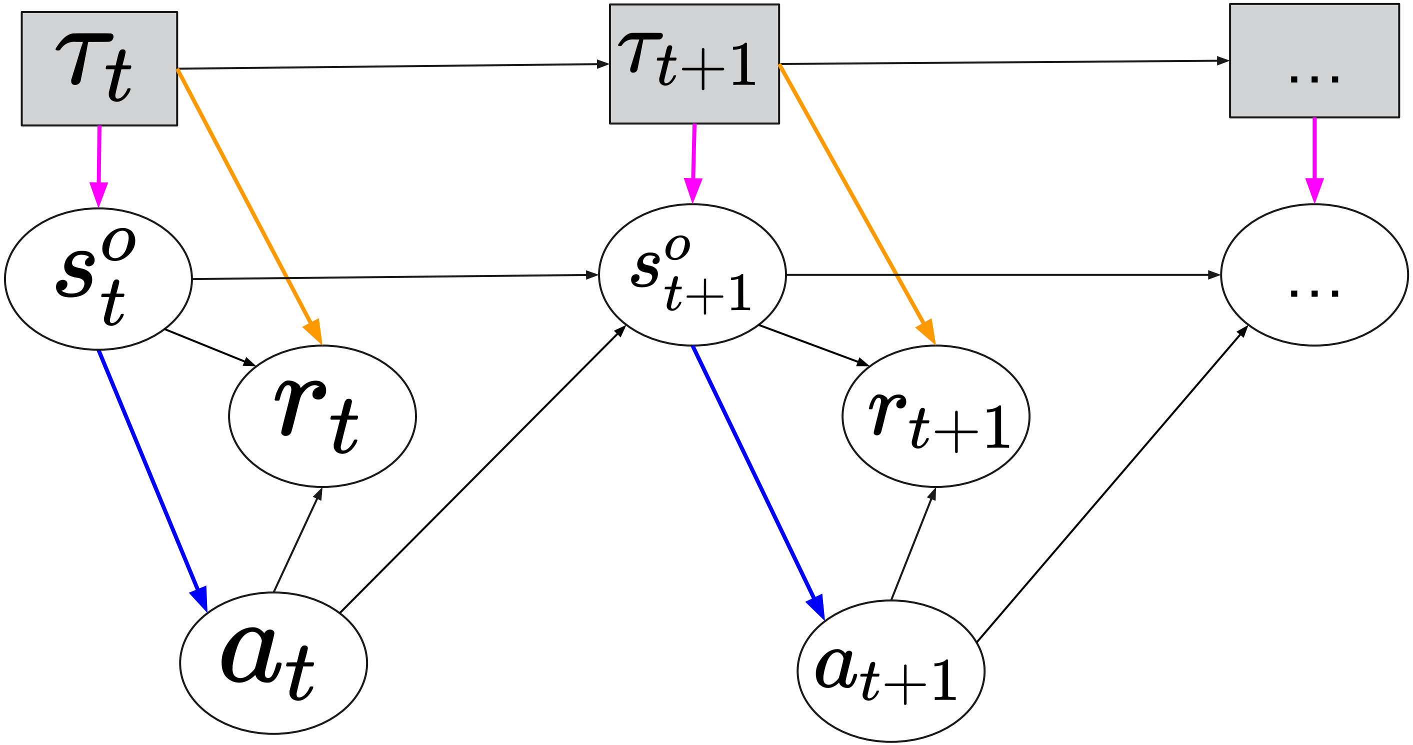
TACRL as a POMDP special case



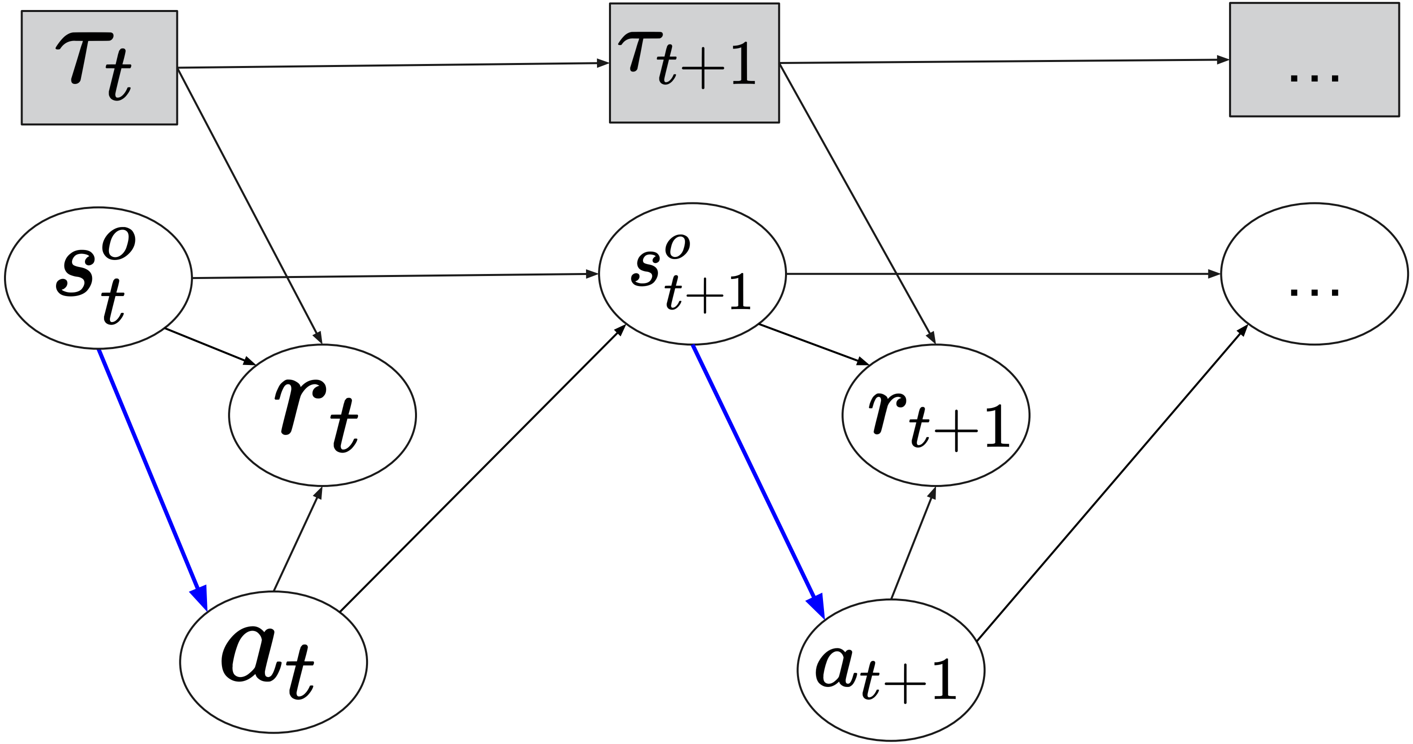
TACRL as a POMDP special case



Task-dependant reward function and dynamics



TACRL (in our case)



Problem

POMDP

\+ locally-stationary hidden states

\+ Passive non-stationarity

\+ single hidden state (task)

Task-Agnostic Continual RL

Task-Awareness soft upper Bound

- Provide task label to agent:
  - POMDP $\rightarrow$ MDP

multi-task soft upper Bound

- Train on a stationary data distribution
    - *Catastrophic Forgetting* disappears

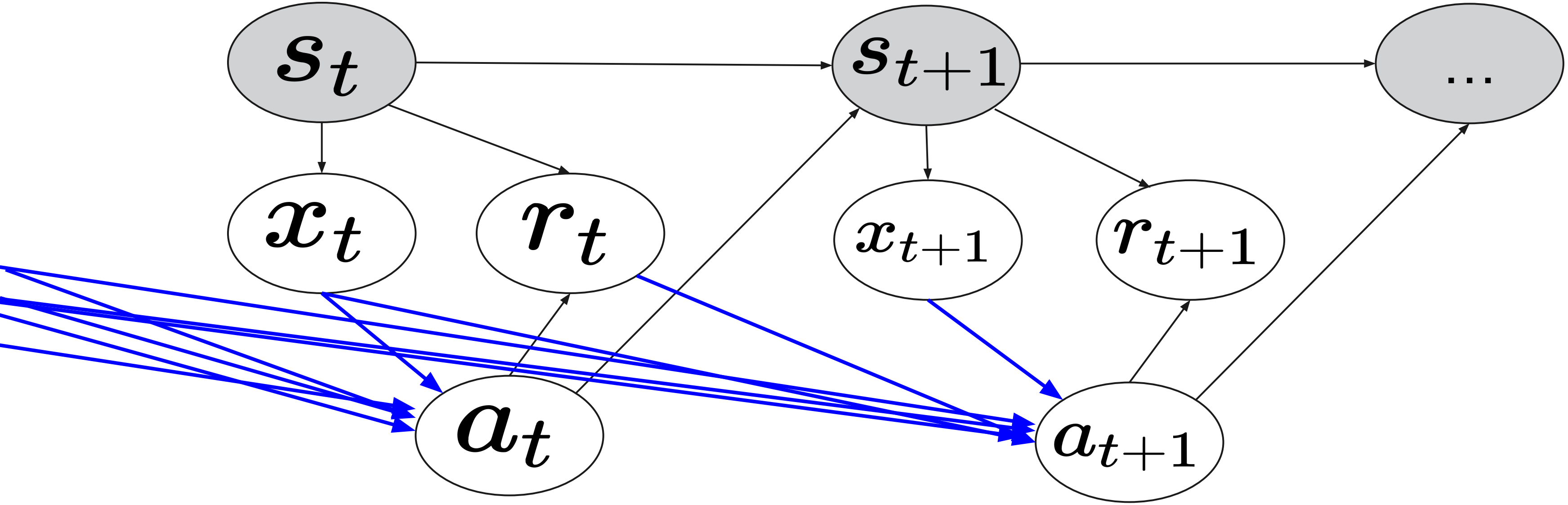
Related settings

	T	π	Objective	Evaluation
MDP [51]	$p(s_{t+1}\|s_t, a_t)$	$\pi(a_t\|s_t)$	$\mathbb{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^t r_t]$	-
POMDP [24]	$p(s^h_{t+1}, s^o_{t+1}\|s^h_t, s^o_t, a_t)$	$\pi(a_t\|s^o_{1:t}, a_{1:t-1}, r_{1:t-1})$	$\mathbb{E}_{s^h}\left[\mathbb{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^t r_t]\|s^h\right]$	-
HM-MDP [9]	$p(s^o_{t+1}\|s^h_{t+1}, s^o_t, a_t)p(s^h_{t+1}\|s^h_t)$	$\pi(a_t\|s^o_{1:t}, a_{1:t-1}, r_{1:t-1})$	$\mathbb{E}_{s^h}\left[\mathbb{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^t r_t]\|s^h\right]$	-
Task-agnostic CRL	$p(s^o_{t+1}\|s^h_{t+1}, s^o_t, a_t)p(s^h_{t+1}\|s^h_t)$	$\pi(a_t\|s^o_{1:t}, a_{1:t-1}, r_{1:t-1})$	$\mathbb{E}_{s^h}\left[\mathbb{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^t r_t]\|s^h\right]$	$\mathbb{E}_{\tilde{s}^h}\left[\mathbb{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^t r_t]\|s^h\right]$
Task-Aware CRL	$p(s^o_{t+1}\|s^h_{t+1}, s^o_t, a_t)p(s^h_{t+1}\|s^h_t)$	$\pi(a_t\|s^h_t, s^o_t)$	$\mathbb{E}_{s^h}\left[\mathbb{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^t r_t]\|s^h\right]$	$\mathbb{E}_{\tilde{s}^h}\left[\mathbb{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^t r_t]\|s^h\right]$
Multi-task RL	$p(s^o_{t+1}\|s^h_{t+1}, s^o_t, a_t)p(s^h_{t+1})$	$\pi(a_t\|s^h_t, s^o_t)$	$\mathbb{E}_{\tilde{s}^h}\left[\mathbb{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^t r_t]\|s^h\right]$	-

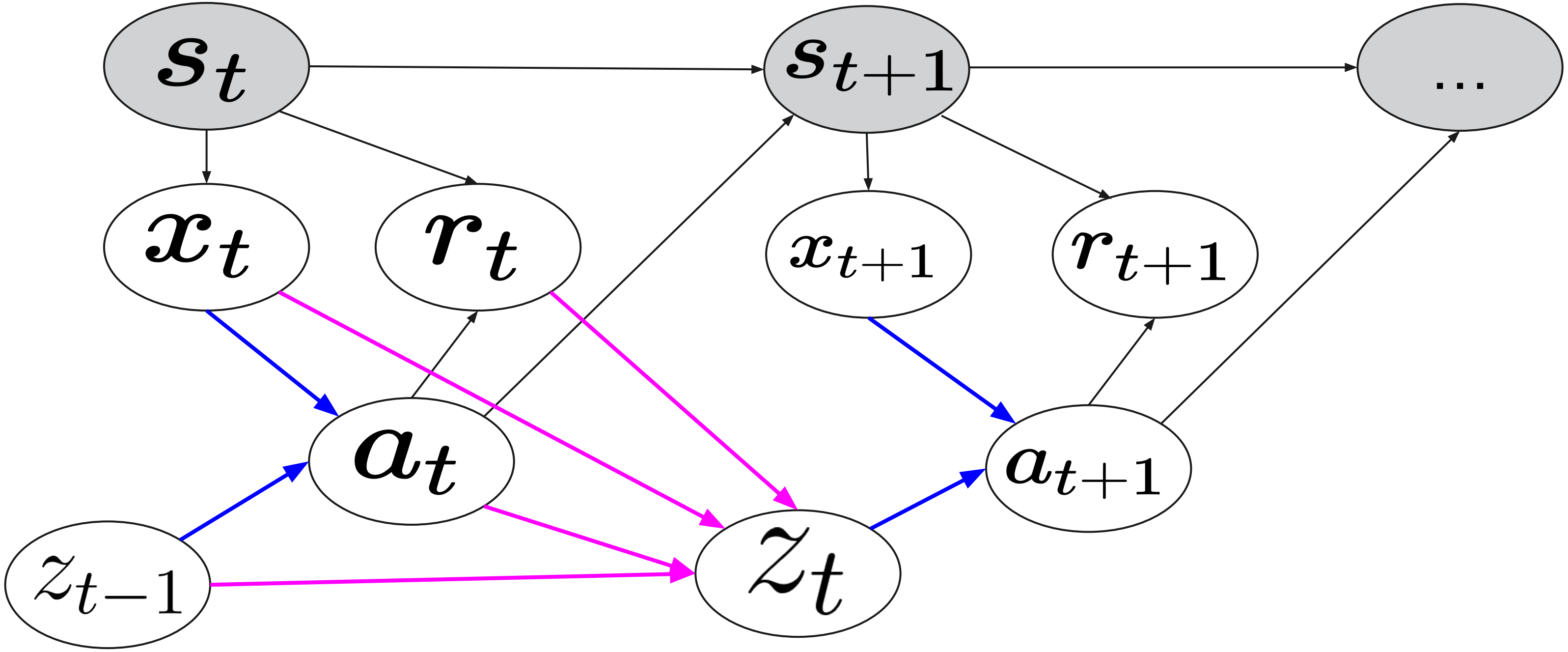
Table 1: Summarizing table of the settings relevant to TACRL. For readability purposes, $\tilde{s}^h$ denotes the stationary distribution of s^h. The Evaluation column if left blank when it is equivalent to the Objective one.

Handling partial observability

Reinforcement Learning (RL): POMDP



Handling partial observability w/ a working memory



Replay-based Recurrent RL (3RL)

- Trick to handle partial observability → Working Memory (RNN)

- Trick to handle forgetting → Experience Replay

- RL algo + RNN + Experience Replay = 

Other baselines

- Strategies
  - Fine Tuning
  - Experience Replay
  - Multi-task

- Task-Aware Modeling:

  += Multi-head (MH)

  +=TaskID

SAC

Q-learning: $Q(s,a) \quad \pi(s) = \mathrm{argmax}_a Q(s,a)$

Doesn't scale!

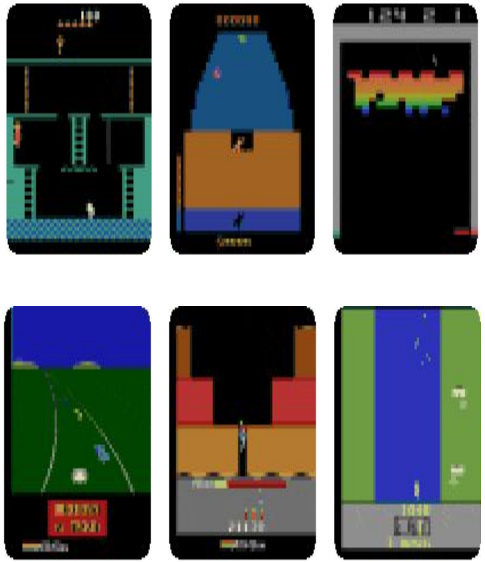
Deep Q-learning (DQN) $Q_\theta(s,a) \quad \pi(s) = \mathrm{argmax}_a Q_\theta(s,a)$

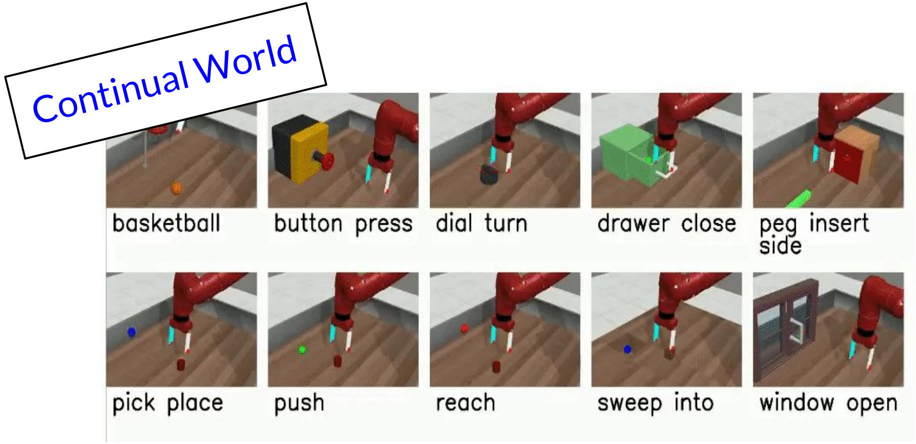
What if my actions are continuous??

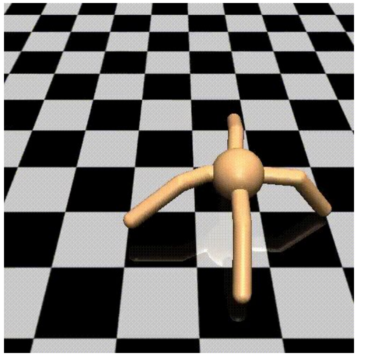
Soft-Actor critic (SAC) $\pi_\phi(s) \approx \mathrm{argmax}_\phi Q_\theta(s, \pi_\phi(s))$

Empirical Findings

Benchmark

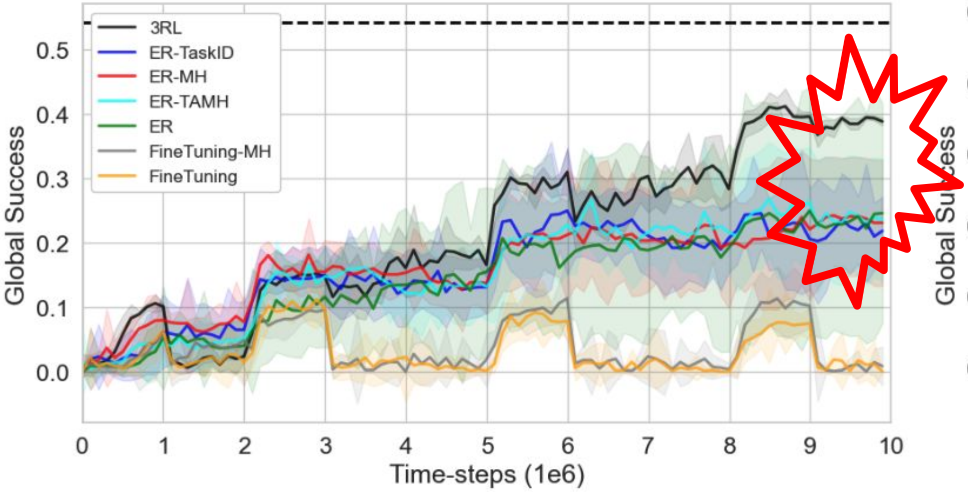


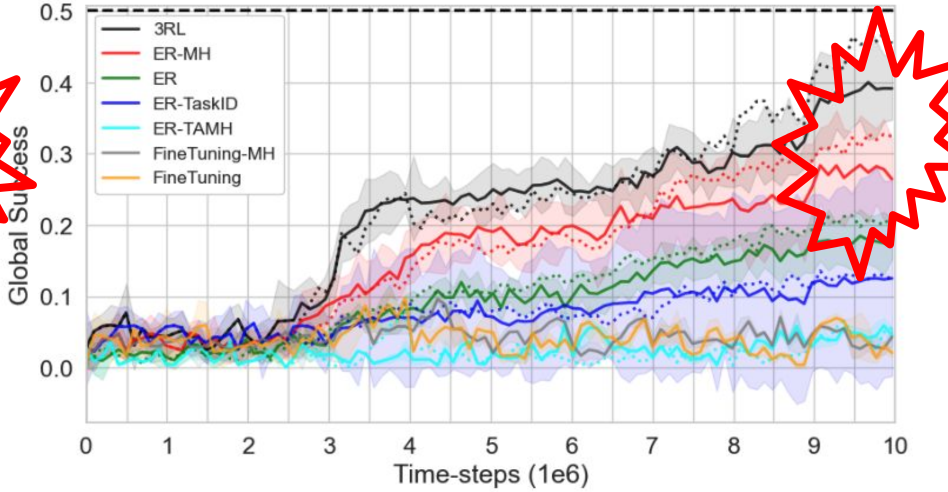




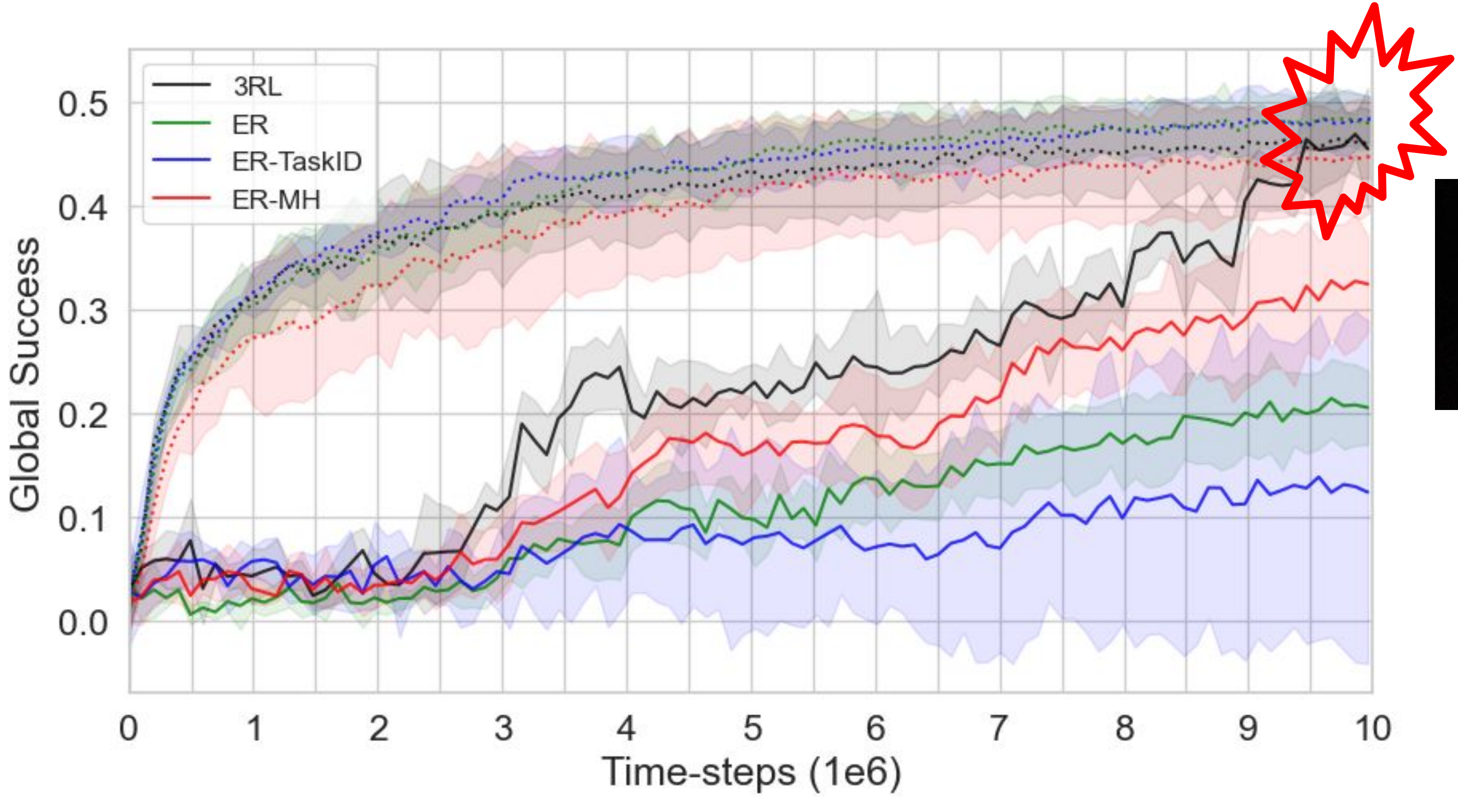
Task similarity

Task-agnostic > Task-Aware



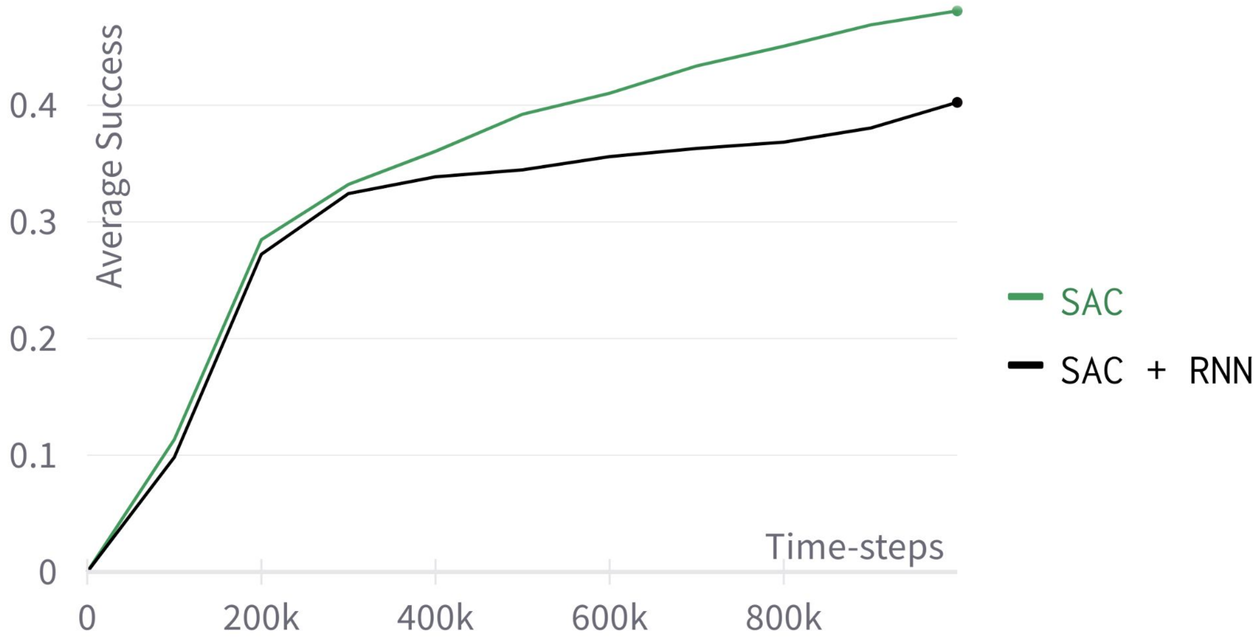


CL = MTL

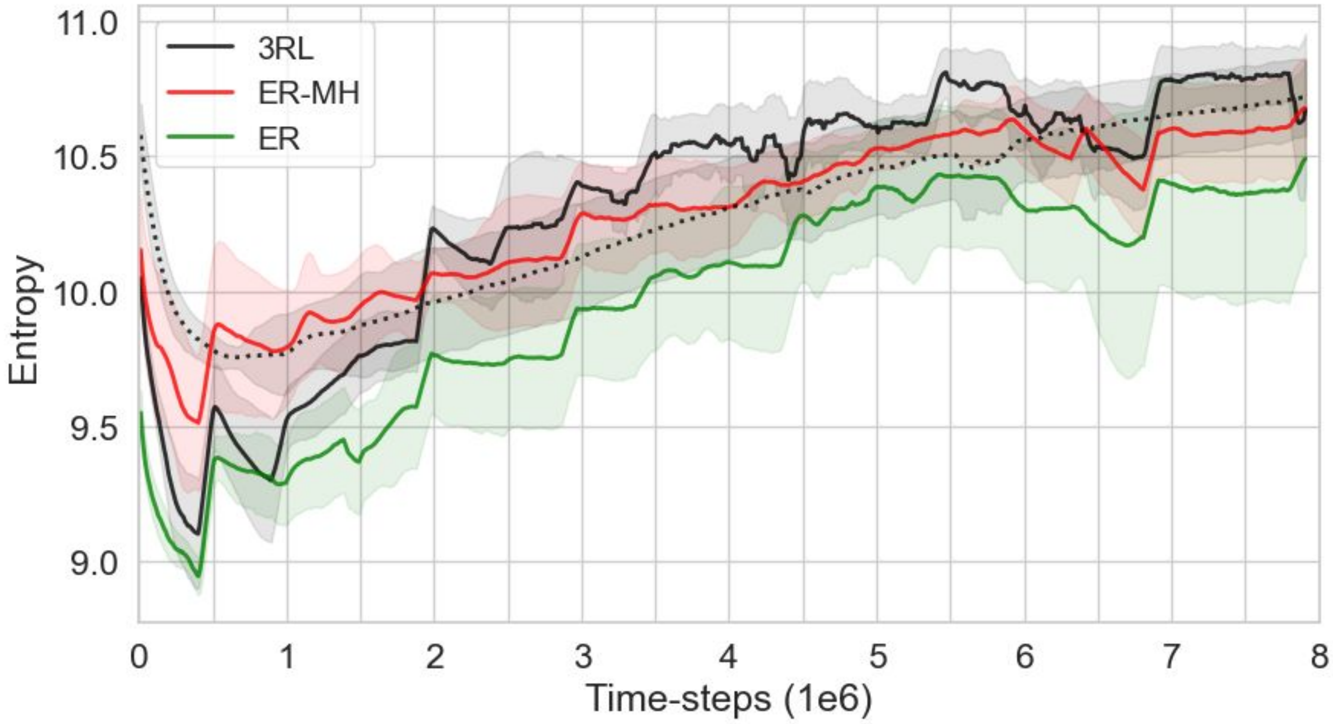

3RL
ER
ER-TaskID
ER-MH
Global Success
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
7
8
9
10
Time-steps (1e6)

Empirical Findings

Hypothesis #1: rnn individually improves single-task performance

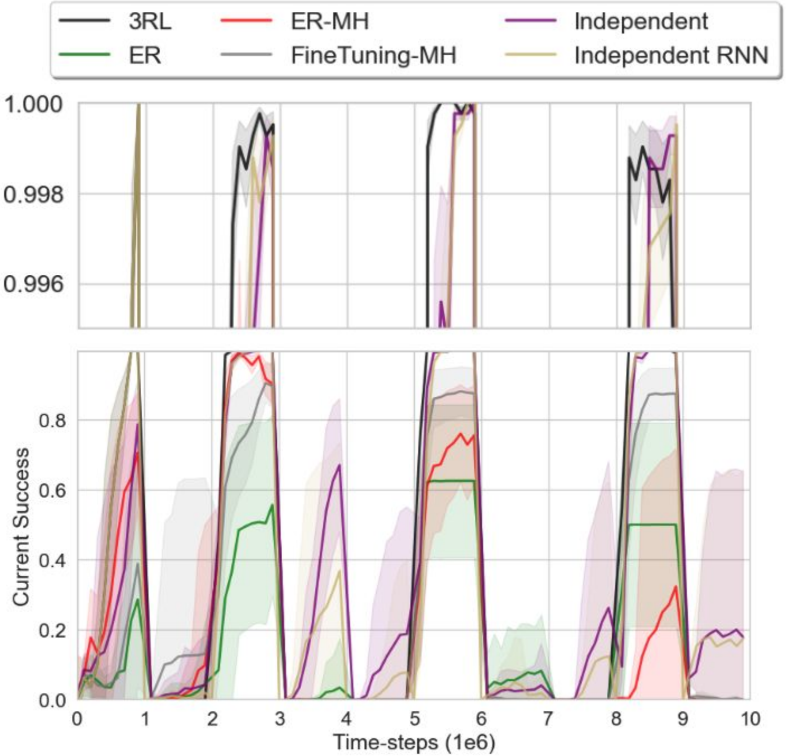


Hypothesis #2: increases parameter stability



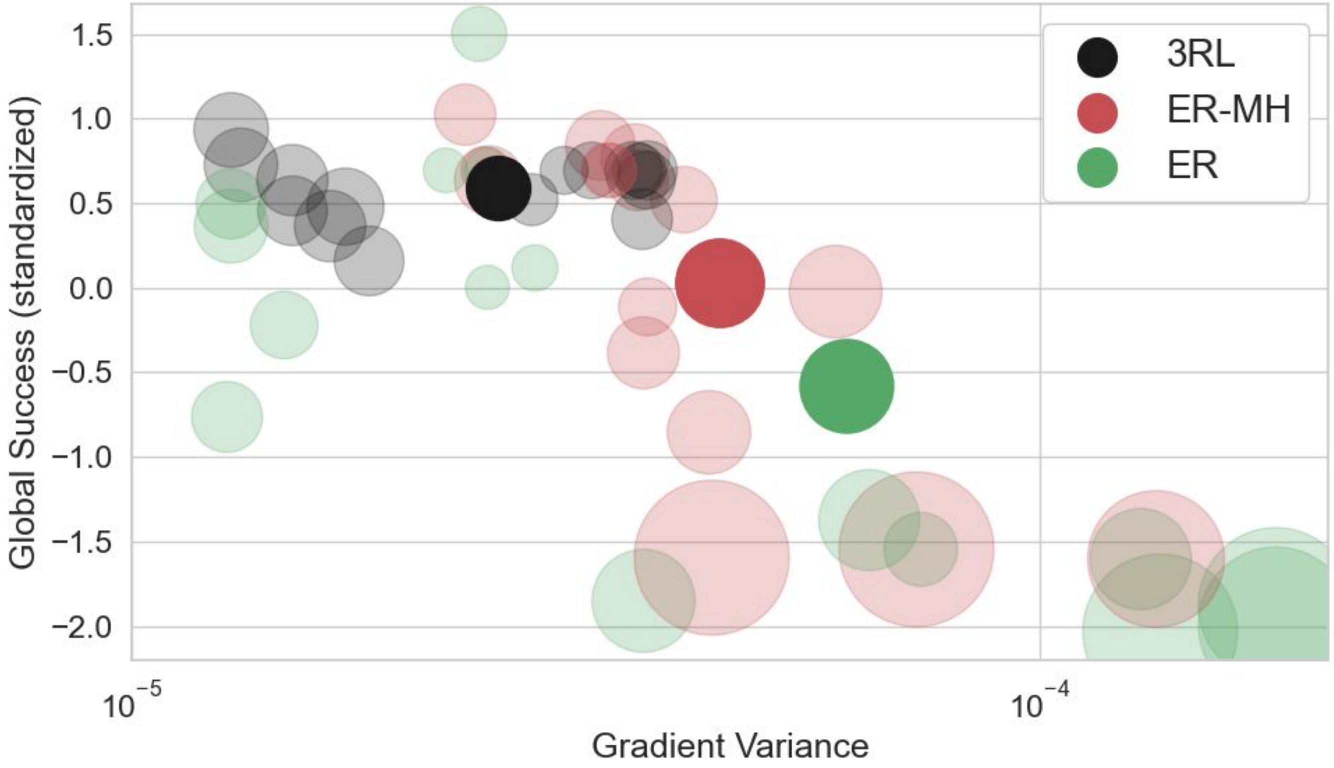
Empirical Findings

Hypothesis #3: rnn correctly places the new tasks in the context for previous ones

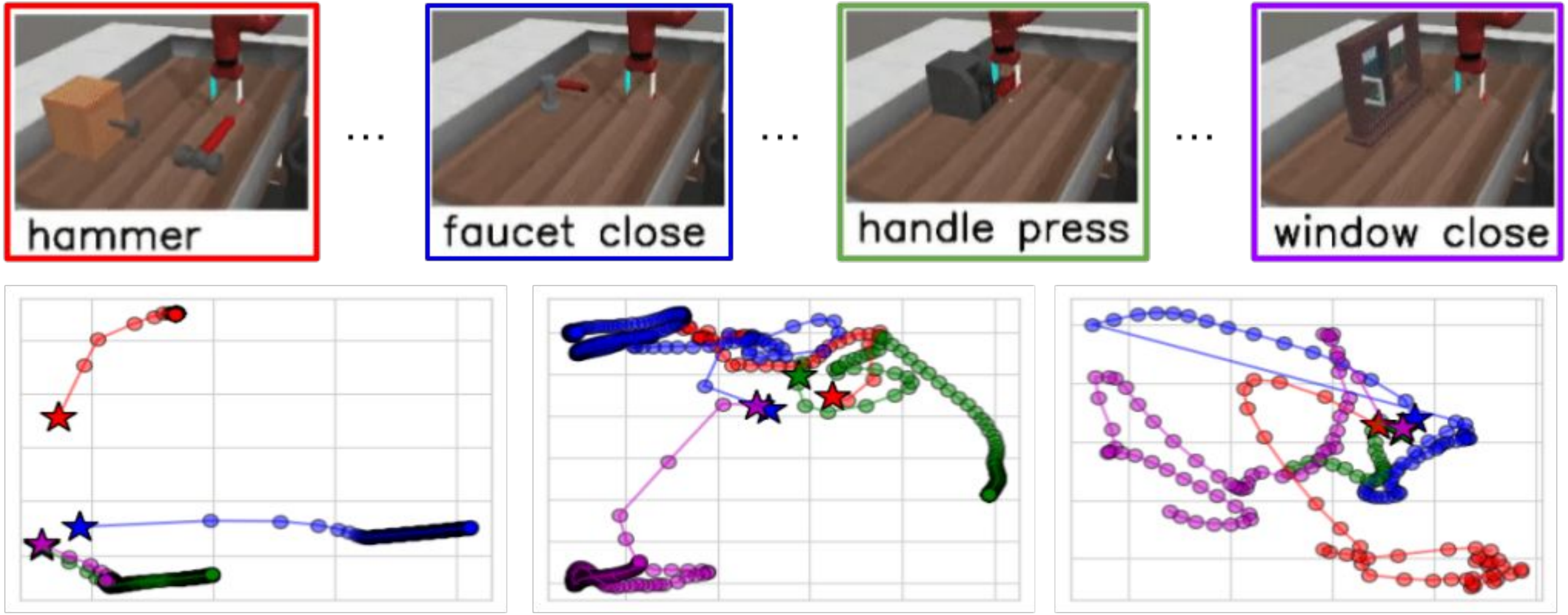


Empirical Findings

Hypothesis #3: rnn correctly places the new tasks in the context for previous ones



Qualitative analysis of the representations



Before training

After one task

End of training

Discussion

- Implications for CL research

- CSL vs CRL