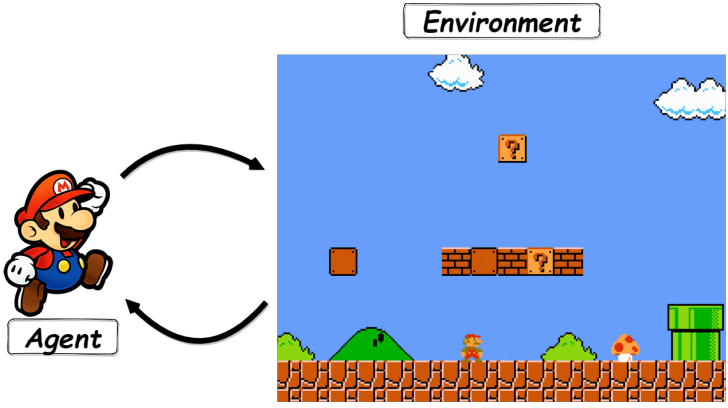
Generalized Hindsight for Reinforcement Learning

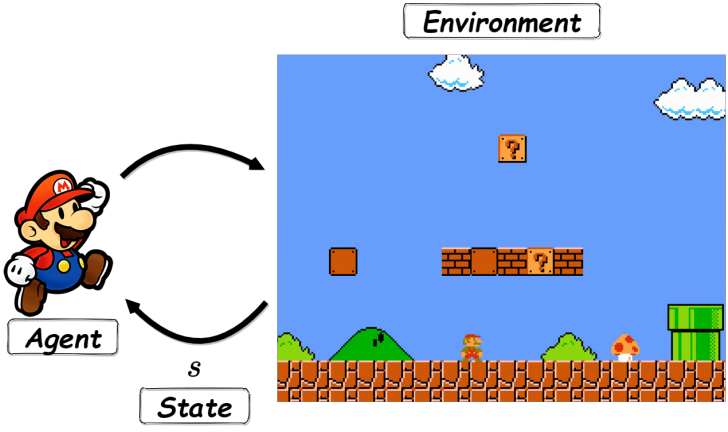
Paper from NeurIPS 2020 by
Alexander C. Li Lerrel Pinto Pieter Abbeel
https://arxiv.org/pdf/2002.11708.pdf

October 20, 2020

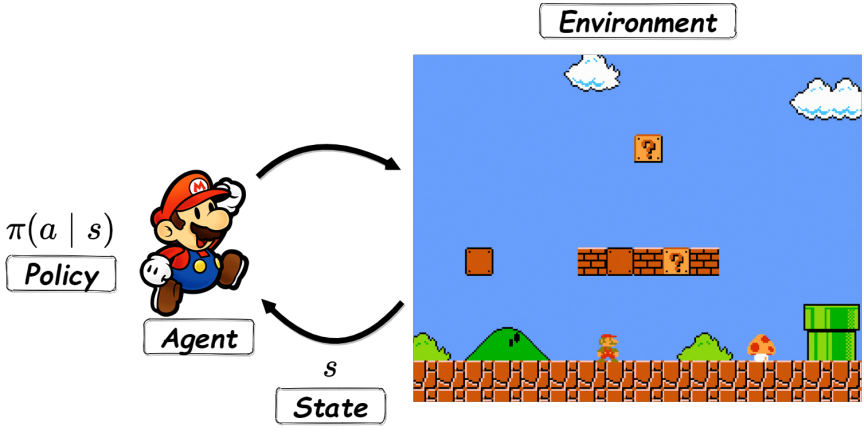
Reinforcement learning



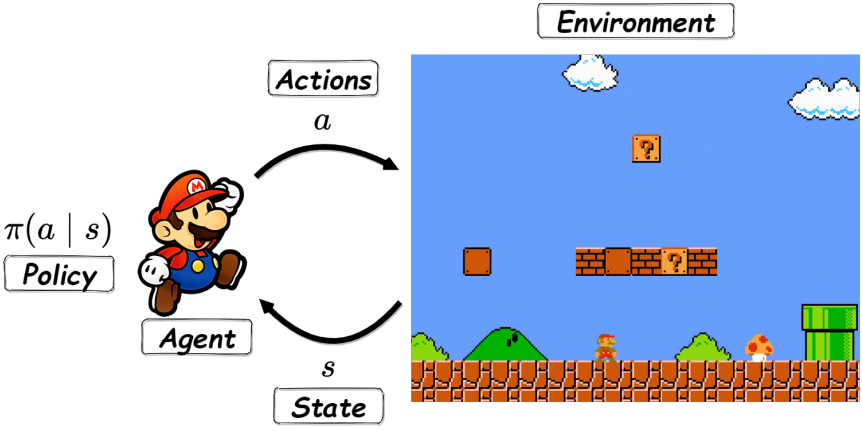
Reinforcement learning



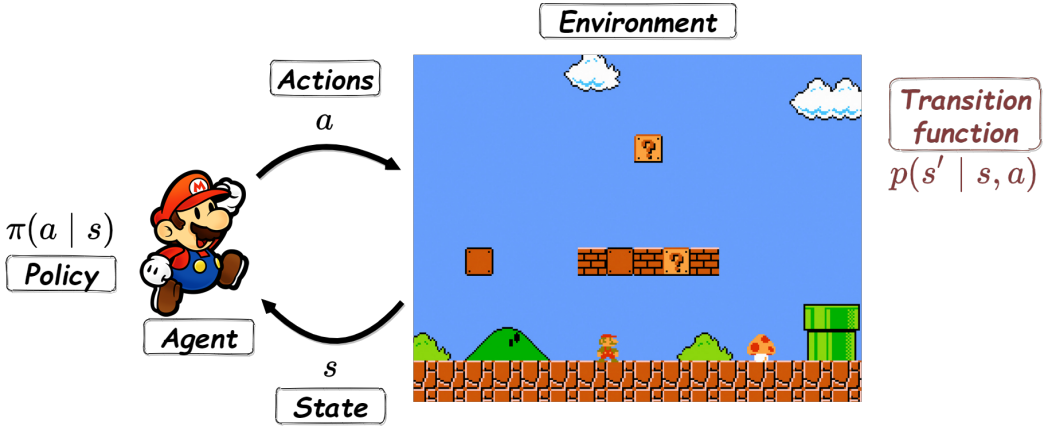
Reinforcement learning



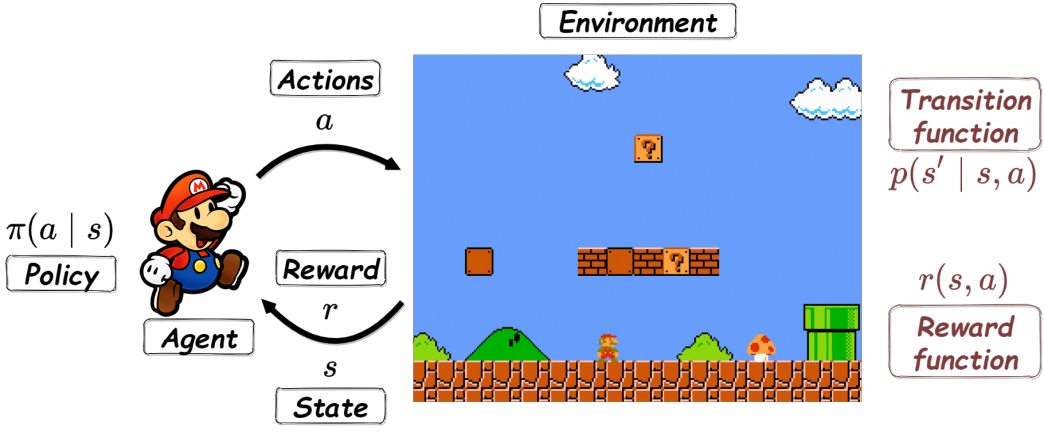
Reinforcement learning



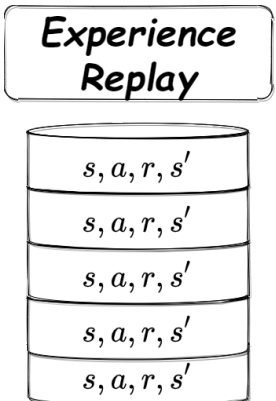
Reinforcement learning



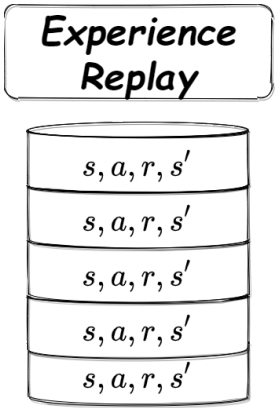
Reinforcement learning



Experience Replay



Experience Replay

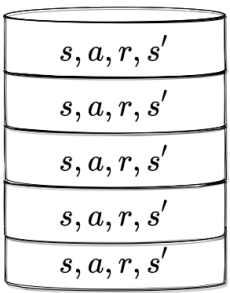


Key property:

$s' \sim p(s' \mid s, a)$

Experience Replay

Experience Replay



Key property:

$s' \sim p(s' \mid s, a)$

Off-policy algorithms
can train from arbitrary replay buffer.

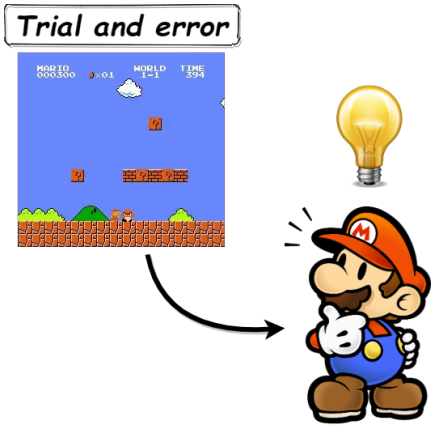
s, a, r, s'

Transition

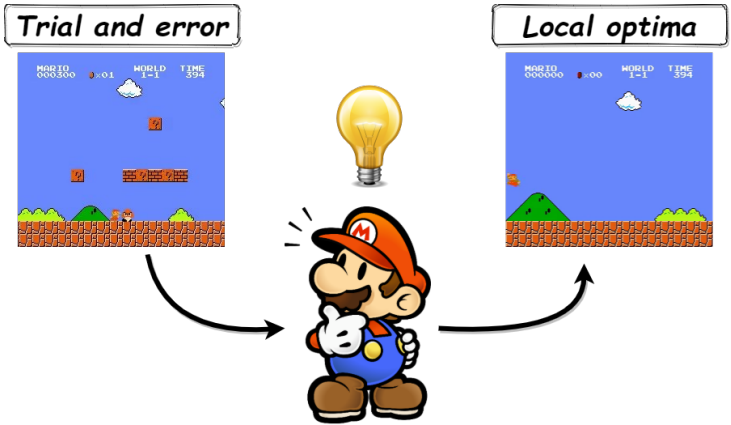
Trial and Error learning

Trial and error

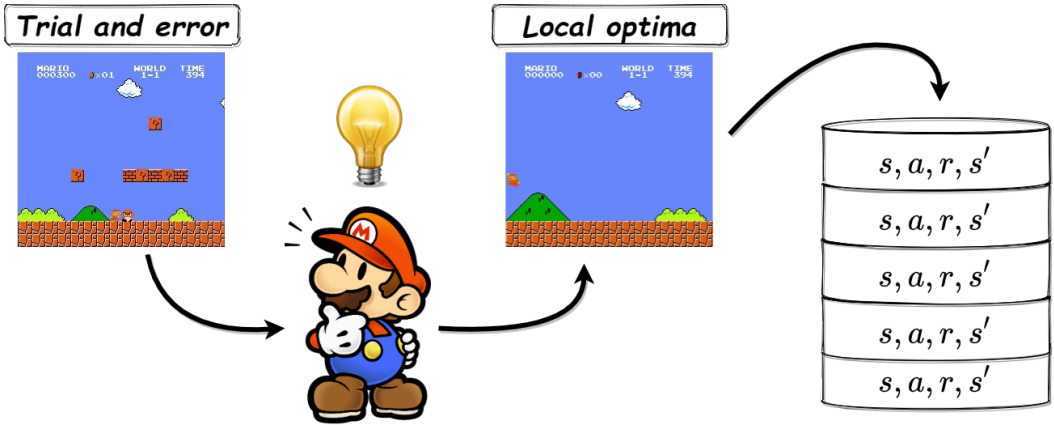
Trial and Error learning



Trial and Error learning



Trial and Error learning



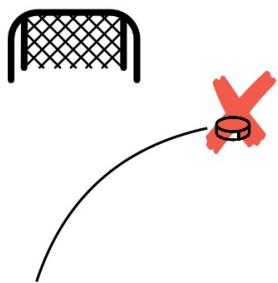
The Worst Case Scenario

Sparse reward setting:
*(sometimes — **search** task)*

$$r(s,a) = \begin{cases} +1 & s \in \mathcal{S}^+ \\ \text{const} & s \notin \mathcal{S}^+ \end{cases}$$

[1] picture source

The Worst Case Scenario



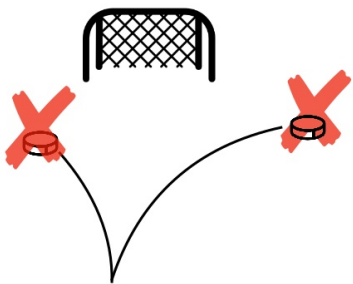
Sparse reward setting:
*(sometimes — **search** task)*

$$r(s,a) = \begin{cases} +1 & s \in \mathcal{S}^+ \\ \text{const} & s \notin \mathcal{S}^+ \end{cases}$$

- $\mathcal{S}^+$ — set of terminal states
- const ≤ 0 — time loosing penalty

[1] picture source

The Worst Case Scenario



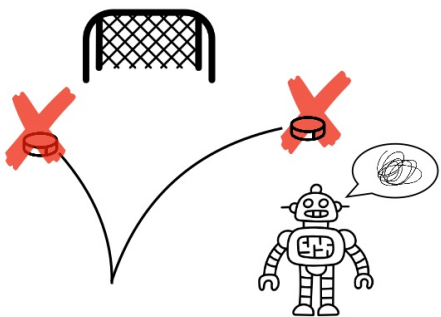
Sparse reward setting:
*(sometimes — **search** task)*

$$r(s,a) = \begin{cases} +1 & s \in \mathcal{S}^+ \\ \text{const} & s \notin \mathcal{S}^+ \end{cases}$$

- $\mathcal{S}^+$ — set of terminal states
- const ≤ 0 — time loosing penalty

[1] picture source

The Worst Case Scenario



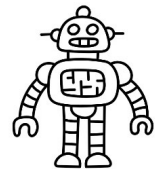
Sparse reward setting:
*(sometimes — **search** task)*

$$r(s,a) = \begin{cases} +1 & s \in \mathcal{S}^+ \\ \text{const} & s \notin \mathcal{S}^+ \end{cases}$$

- $\mathcal{S}^+$ — set of terminal states
- const ≤ 0 — time loosing penalty

[1] picture source

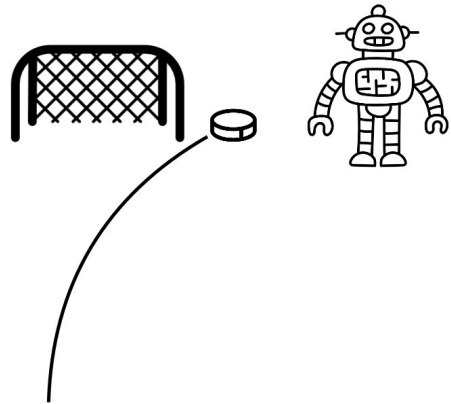
Hindsight: idea



[1] picture source

Hindsight: idea

× no hope at all?

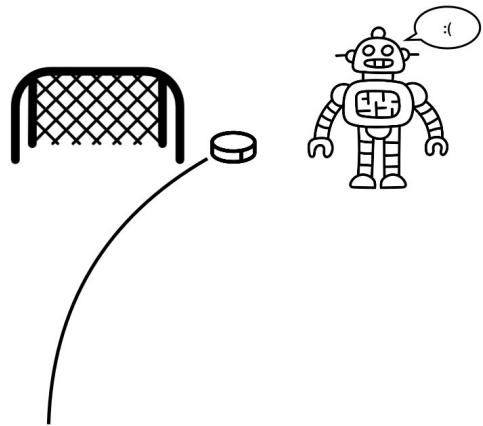


[1] picture source

Hindsight: idea

× no hope at all?

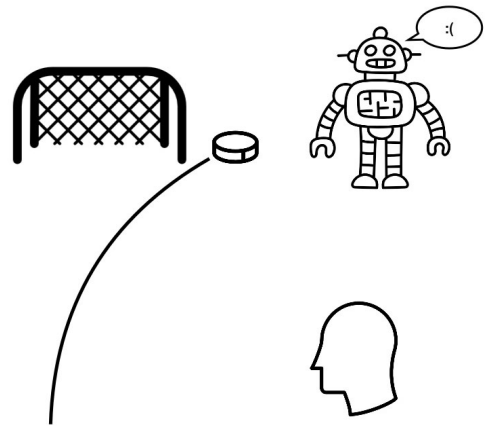
× no reward signal



[1] picture source

Hindsight: idea

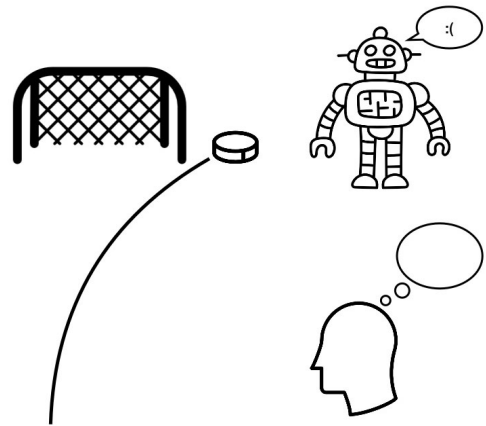
- × no hope at all?
- × no reward signal
- × no training available



[1] picture source

Hindsight: idea

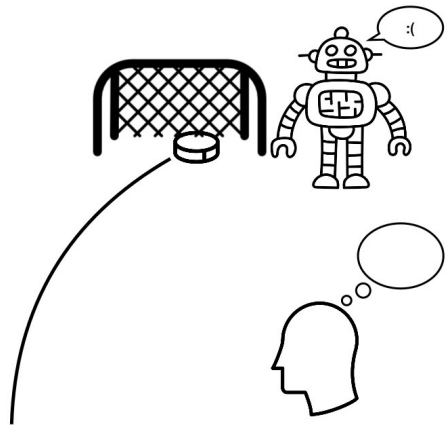
- × no hope at all?
- × no reward signal
- × no training available



[1] picture source

Hindsight: idea

- × no hope at all?
- × no reward signal
- × no training available



[1] picture source

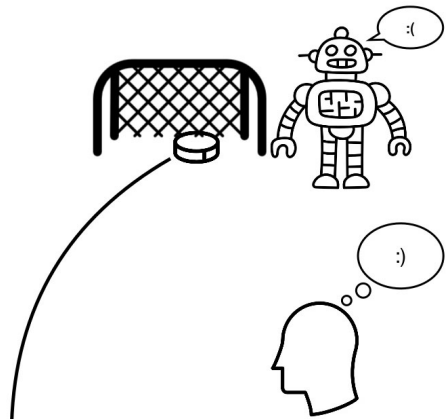
Hindsight: idea

× no hope at all?

× no reward signal

× no training available

If last $s \notin \mathcal{S}^+$, you can **pretend** that it was («*in hindsight*»)!



[1] picture source

Hindsight: idea

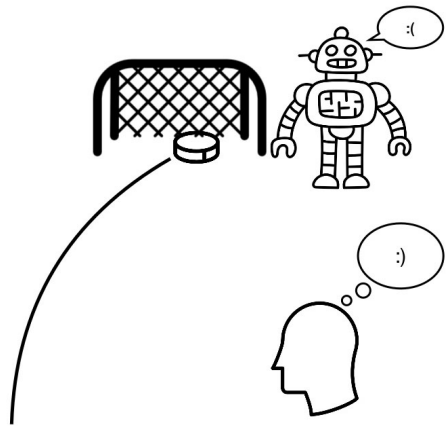
× no hope at all?

× no reward signal

× no training available

If last $s \notin \mathcal{S}^+$, you can **pretend** that it was («*in hindsight*»)!

	Data	**Goal**
Previously:	$(s, a, +0, s')$	$\mathcal{S}^+$



[1] picture source

Hindsight: idea

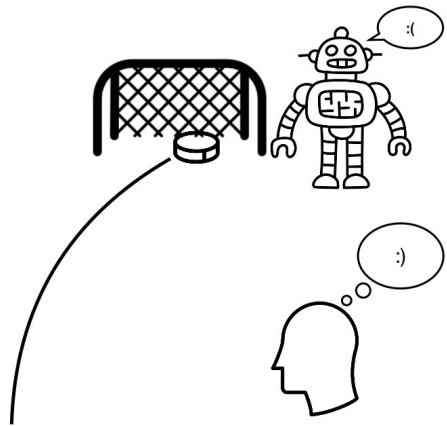
× no hope at all?

× no reward signal

× no training available

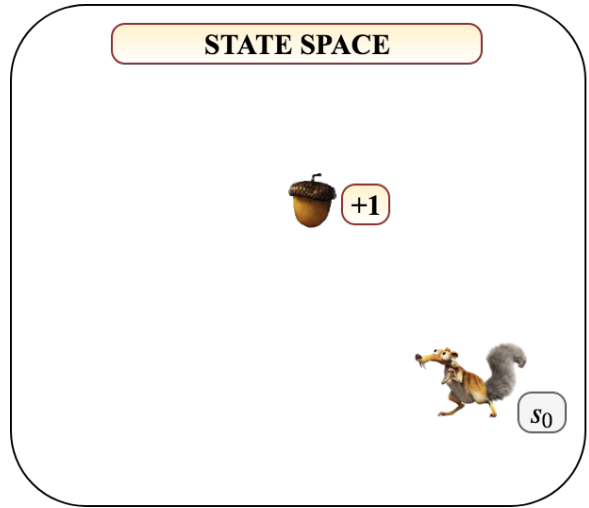
If last $s \notin \mathcal{S}^+$, you can **pretend** that it was («*in hindsight*»)!

	Data	**Goal**
Previously:	$(s, a, +0, s')$	$\mathcal{S}^+$
Now:	$(s, a, +1, s')$	$\{s'\}$



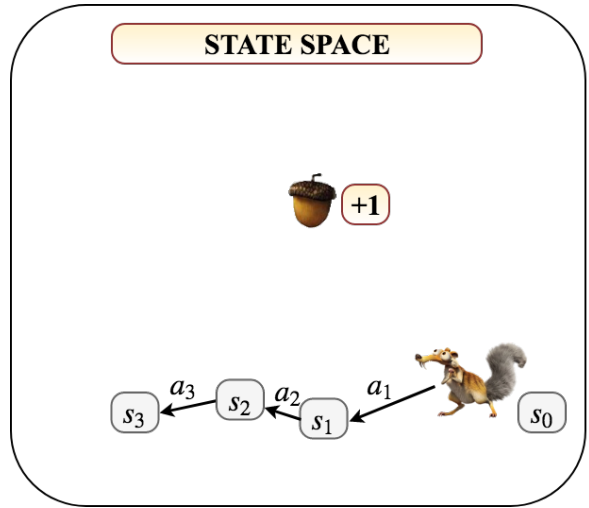
[1] picture source

Hindsight Experience Replay[1] (2017)



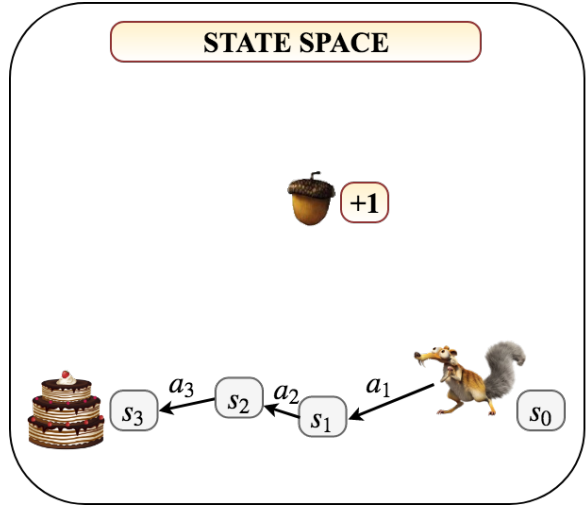
[1] https://arxiv.org/abs/1707.01495

Hindsight Experience Replay[1] (2017)



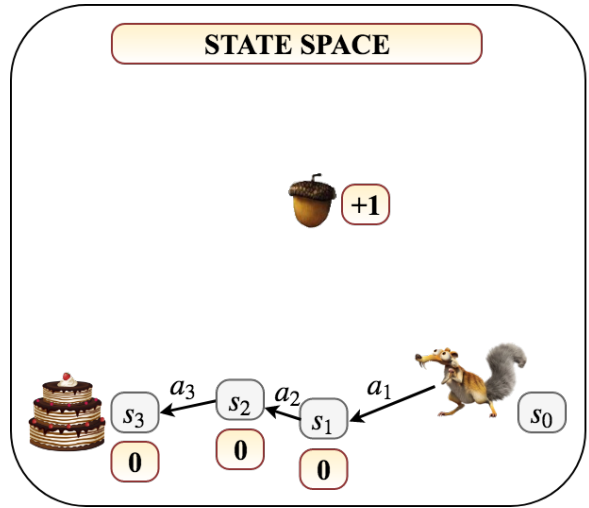
[1] https://arxiv.org/abs/1707.01495

Hindsight Experience Replay[1] (2017)



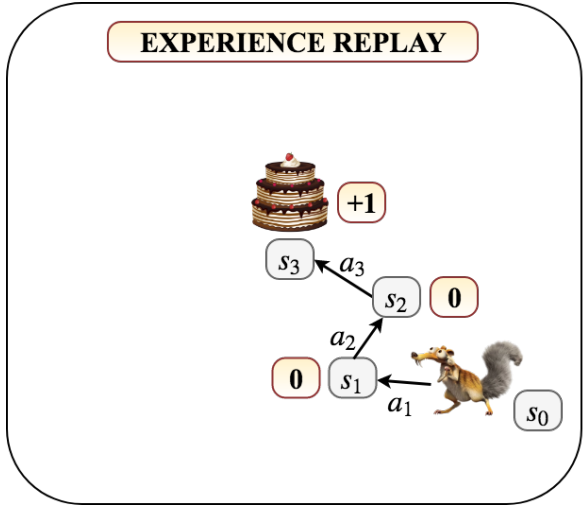
[1] https://arxiv.org/abs/1707.01495

Hindsight Experience Replay[1] (2017)



[1] https://arxiv.org/abs/1707.01495

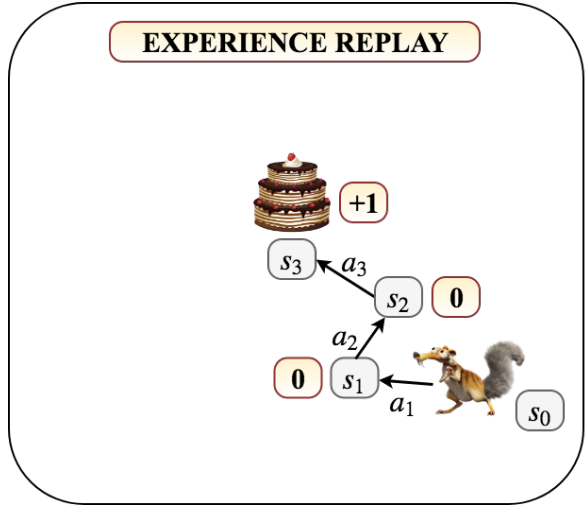
Hindsight Experience Replay[1] (2017)



[1] https://arxiv.org/abs/1707.01495

Hindsight Experience Replay[1] (2017)

We performed **trajectory relabeling**.



[1] https://arxiv.org/abs/1707.01495

HER: why it works

Goal-extended MDP:

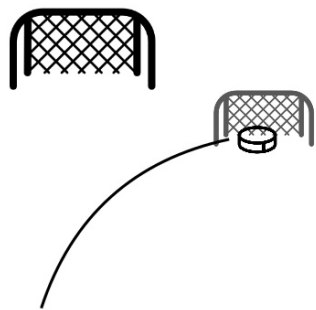
- new state space is $\mathcal{S} \times \mathcal{G}$;
- new reward function is $r(s, g)$;

[1] picture source

HER: why it works

Goal-extended MDP:

- new state space is $\mathcal{S} \times \mathcal{G}$;
- new reward function is $r(s, g)$;



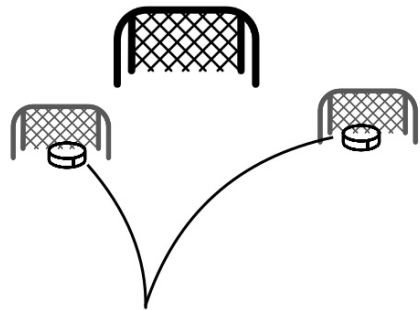
[1] picture source

HER: why it works

Goal-extended MDP:

- new state space is $\mathcal{S} \times \mathcal{G}$;
- new reward function is $r(s, g)$;

We are now training **universal value functions** $Q(s, g, a)$ and **universal policies** $\pi(a \mid s, g)$.



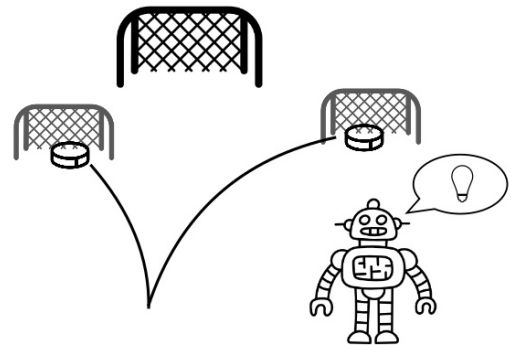
[1] picture source

HER: why it works

Goal-extended MDP:

- new state space is $\mathcal{S} \times \mathcal{G}$;
- new reward function is $r(s, g)$;

We are now training **universal value functions** $Q(s, g, a)$ and **universal policies** $\pi(a \mid s, g)$.



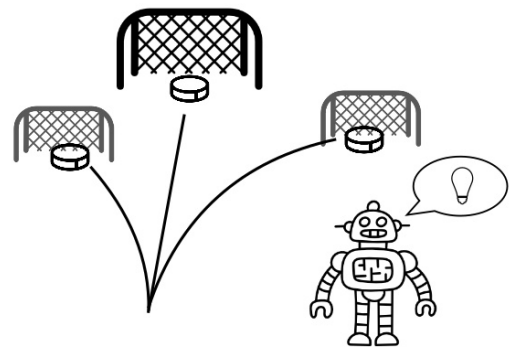
[1] picture source

HER: why it works

Goal-extended MDP:

- new state space is $\mathcal{S} \times \mathcal{G}$;
- new reward function is $r(s, g)$;

We are now training **universal value functions** $Q(s, g, a)$ and **universal policies** $\pi(a \mid s, g)$.



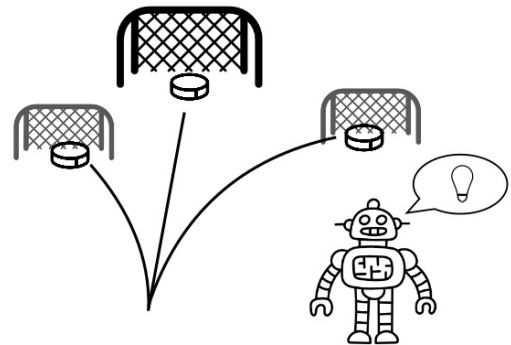
[1] picture source

HER: why it works

Goal-extended MDP:

- new state space is $\mathcal{S} \times \mathcal{G}$;
- new reward function is $r(s, g)$;

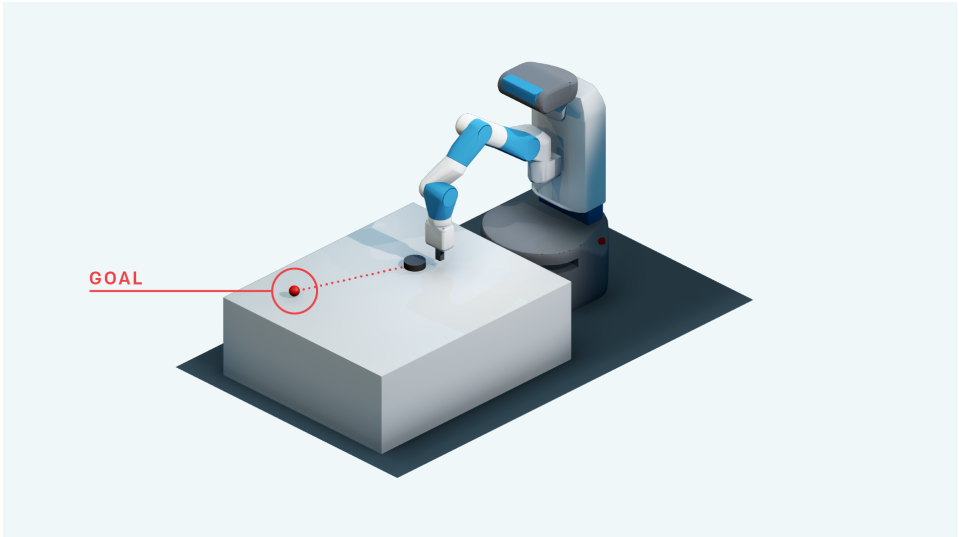
We are now training **universal value functions** $Q(s, g, a)$ and **universal policies** $\pi(a \mid s, g)$.



Case $\mathcal{S} \equiv \mathcal{G}$: learning **navigation** in state space!

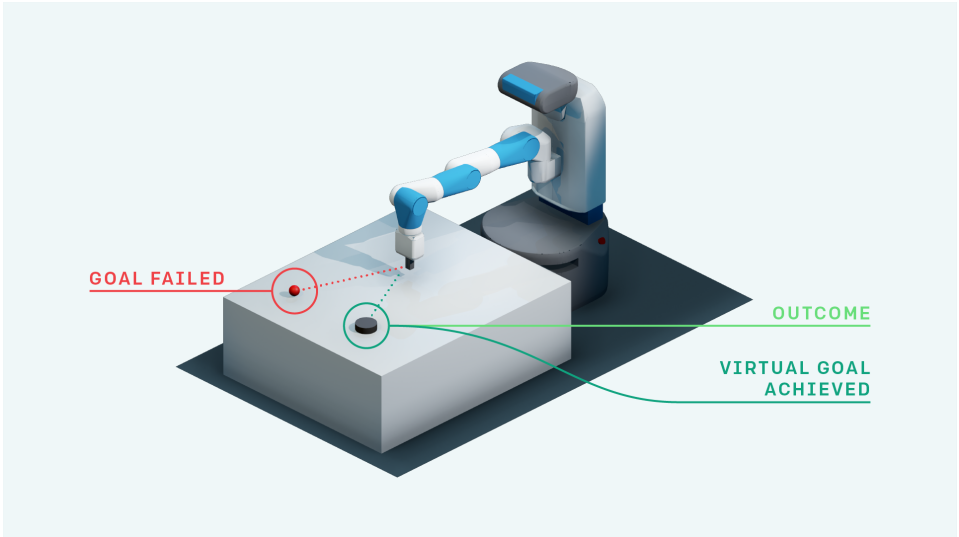
[1] picture source

Standard tasks for hindsight testing



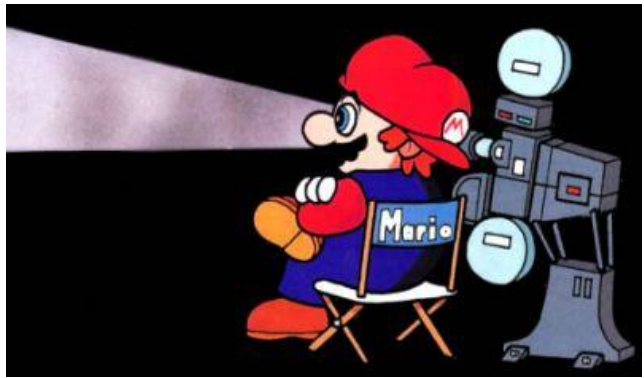
[1] picture source

Standard tasks for hindsight testing



[1] picture source

HER: results[2]



[2] https://www.youtube.com/watch?v=Dz_HuzgMxzo

Multi-task Reinforcement Learning

In arbitrary MDP with goals $\mathcal{S}, \mathcal{G}, \mathcal{A}, \mathcal{P}, r(s, g)$:

- At the beginning of each training episode: $g \sim p(g)$
  - Goal does not change during episode: $g' = g$

Multi-task Reinforcement Learning

In arbitrary MDP with goals $\mathcal{S}, \mathcal{G}, \mathcal{A}, \mathcal{P}, r(s, g)$:

- At the beginning of each training episode: $g \sim p(g)$
  - Goal does not change during episode: $g' = g$

Laws of physics $\mathcal{P}$ are not affected by goals:

$$p(s' \mid s, g, a) = p(s' \mid s, a)$$

Multi-task Reinforcement Learning

In arbitrary MDP with goals $\mathcal{S}, \mathcal{G}, \mathcal{A}, \mathcal{P}, r(s, g)$:

- At the beginning of each training episode: $g \sim p(g)$
  - Goal does not change during episode: $g' = g$

Laws of physics $\mathcal{P}$ are not affected by goals:

$$p(s' \mid s, g, a) = p(s' \mid s, a)$$

Consequence: if (s, g, a, r, s') is valid transition, then $\forall \hat{g} \in \mathcal{G}$: $(s, \hat{g}, a,$ $)$ is valid transition

Multi-task Reinforcement Learning

In arbitrary MDP with goals $\mathcal{S}, \mathcal{G}, \mathcal{A}, \mathcal{P}, r(s, g)$:

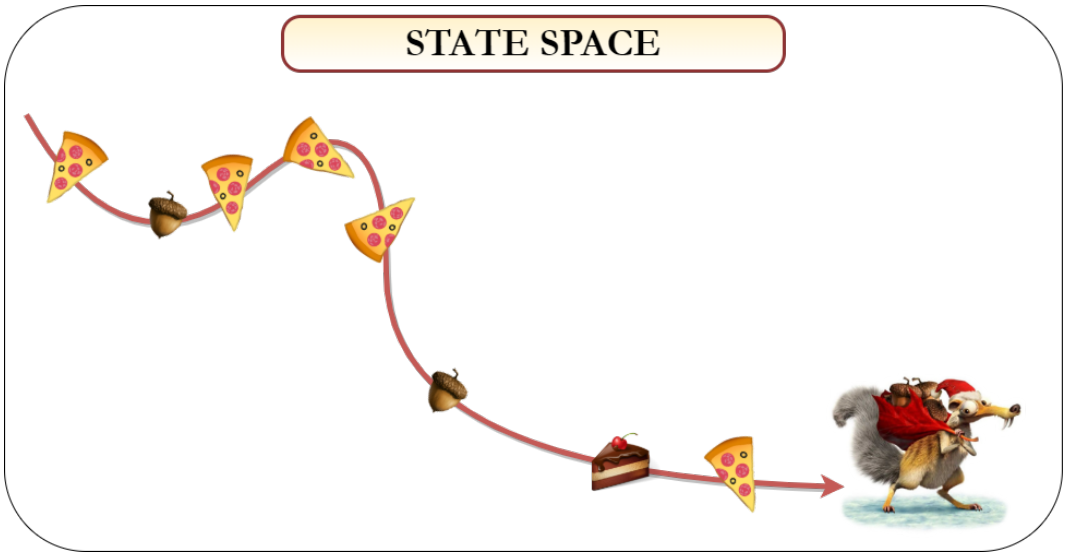
- At the beginning of each training episode: $g \sim p(g)$
  - Goal does not change during episode: $g' = g$

Laws of physics $\mathcal{P}$ are not affected by goals:

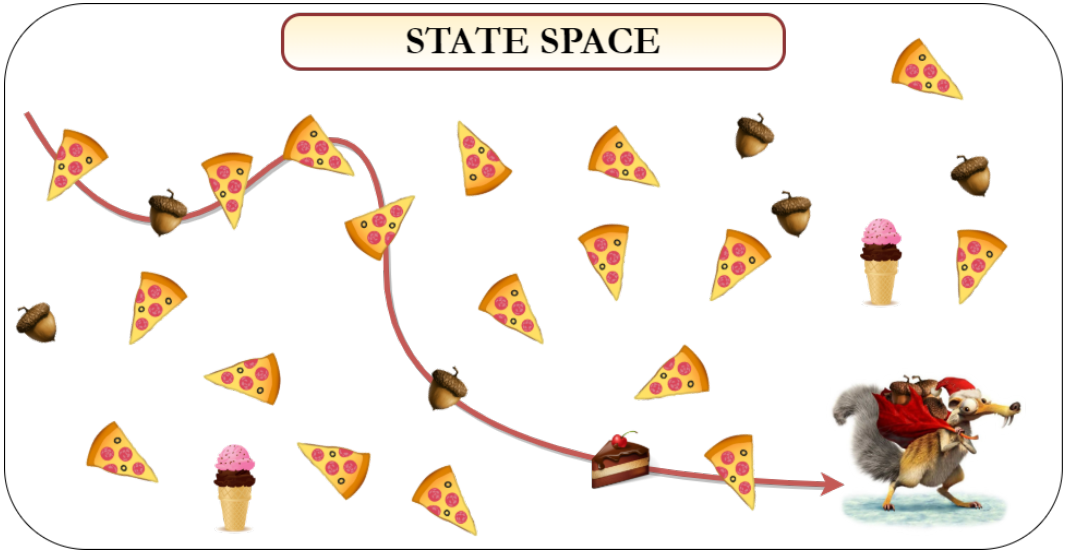
$$p(s' \mid s, g, a) = p(s' \mid s, a)$$

Consequence: if (s, g, a, r, s') is valid transition,
then $\forall \hat{g} \in \mathcal{G}$: $(s, \hat{g}, a, r(s, \hat{g}), s')$ is valid transition
$\Rightarrow$ **knowledge transfer across tasks**.

How to relabel this trajectory?



How to relabel this trajectory?



Inverse Reinforcement Learning (IRL)

- **Given**: expert demonstrations (trajectories) $\mathcal{T}_1, \mathcal{T}_2 \ldots \mathcal{T}_M$ of how to solve some task

[3] Rewriting History with Inverse RL: Hindsight Inference for Policy Improvement, https://arxiv.org/abs/2002.11089

Inverse Reinforcement Learning (IRL)

- **Given**: expert demonstrations (trajectories) $\mathcal{T}_1, \mathcal{T}_2 \ldots \mathcal{T}_M$ of how to solve some task
  - **Find**: reward function, describing this task.

[3] Rewriting History with Inverse RL: Hindsight Inference for Policy Improvement, https://arxiv.org/abs/2002.11089

Inverse Reinforcement Learning (IRL)

- **Given**: expert demonstrations (trajectories) $\mathcal{T}_1, \mathcal{T}_2 \ldots \mathcal{T}_M$ of how to solve some task
- **Find**: reward function, describing this task.

For which g is trajectory $\mathcal{T}$ a sample from optimal behavior aimed at maximizing cumulative $r(s, g)$?

[3] Rewriting History with Inverse RL: Hindsight Inference for Policy Improvement, https://arxiv.org/abs/2002.11089

Inverse Reinforcement Learning (IRL)

- **Given**: expert demonstrations (trajectories) $\mathcal{T}_1, \mathcal{T}_2 \ldots \mathcal{T}_M$ of how to solve some task
- **Find**: reward function, describing this task.

For which g is trajectory $\mathcal{T}$ a sample from optimal behavior aimed at maximizing cumulative $r(s, g)$?

IRL theory[3]:
in concurrent article.

[3] Rewriting History with Inverse RL: Hindsight Inference for Policy Improvement, https://arxiv.org/abs/2002.11089

Inverse Reinforcement Learning (IRL)

- **Given**: expert demonstrations (trajectories) $\mathcal{T}_1, \mathcal{T}_2 \ldots \mathcal{T}_M$ of how to solve some task
- **Find**: reward function, describing this task.

For which g is trajectory $\mathcal{T}$ a sample from optimal behavior aimed at maximizing cumulative $r(s, g)$?

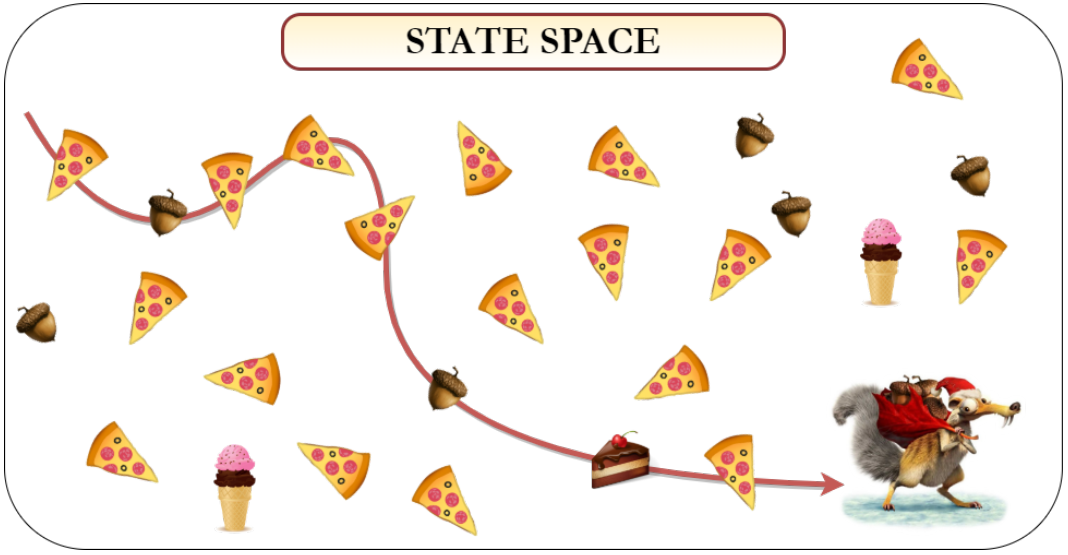
IRL theory[3]:
in concurrent article.

Approximate IRL Relabeling (AIR):

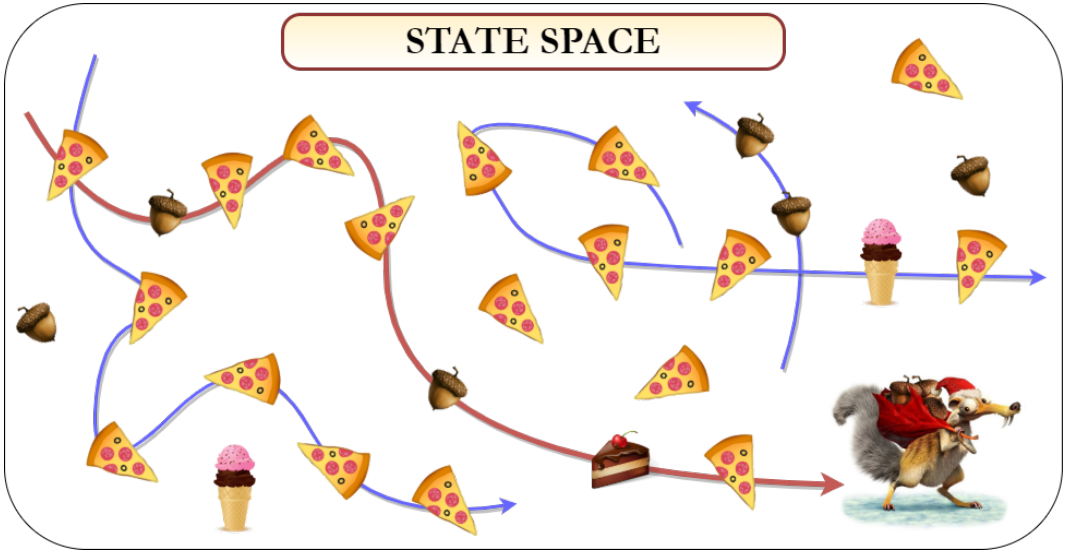
$$\text{approx. } \forall \hat{\mathcal{T}}: \sum_{s \in \mathcal{T}} r(s, g) \geq \sum_{s \in \hat{\mathcal{T}}} r(s, g)$$

[3] Rewriting History with Inverse RL: Hindsight Inference for Policy Improvement, https://arxiv.org/abs/2002.11089

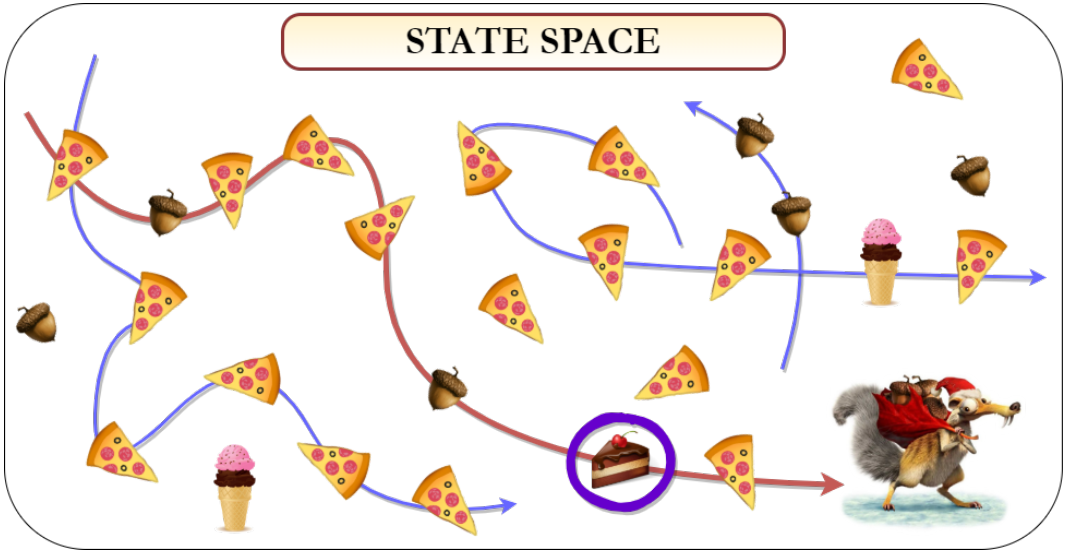
Approximate IRL Relabeling (AIR)



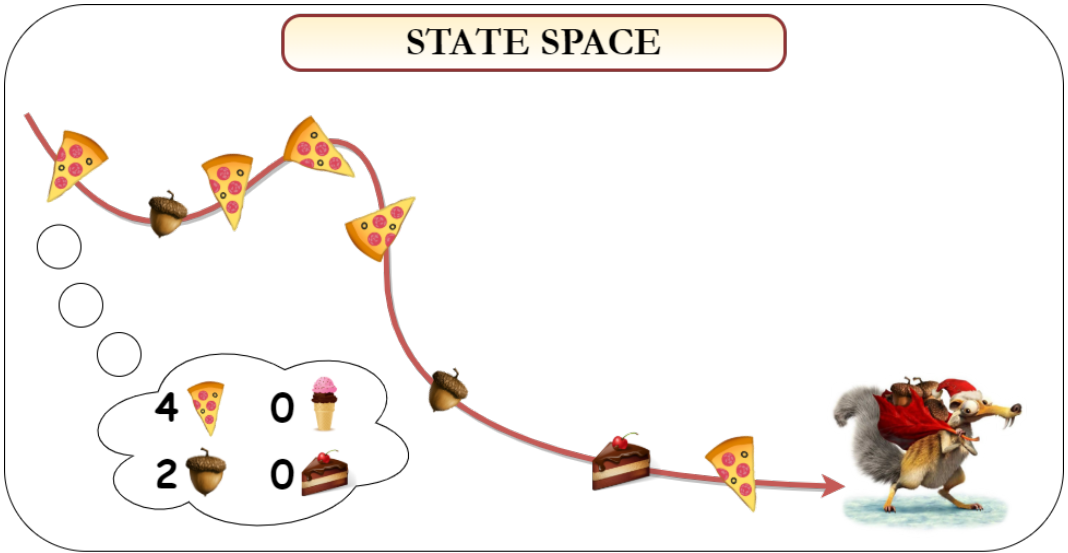
Approximate IRL Relabeling (AIR)



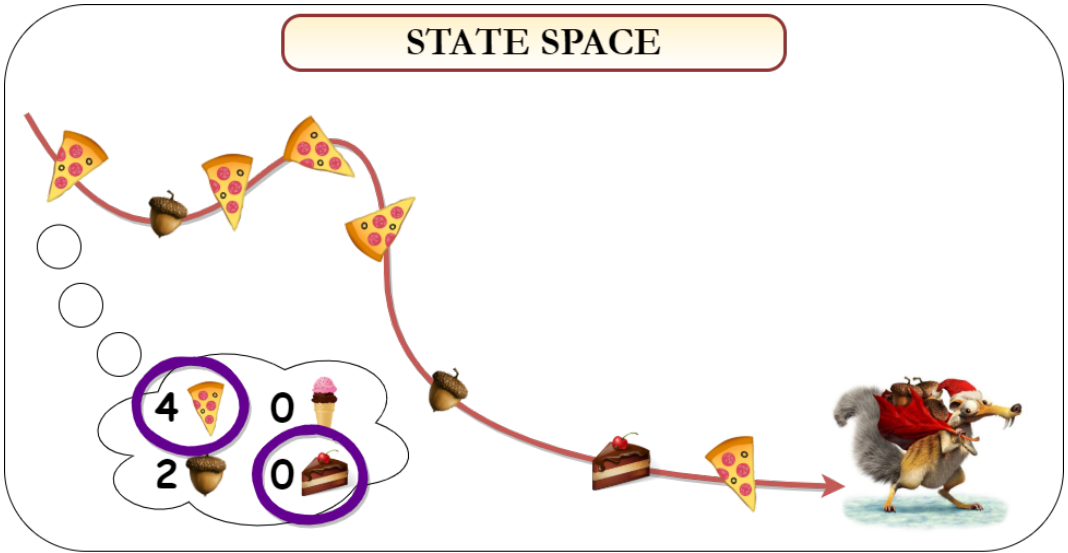
Approximate IRL Relabeling (AIR)



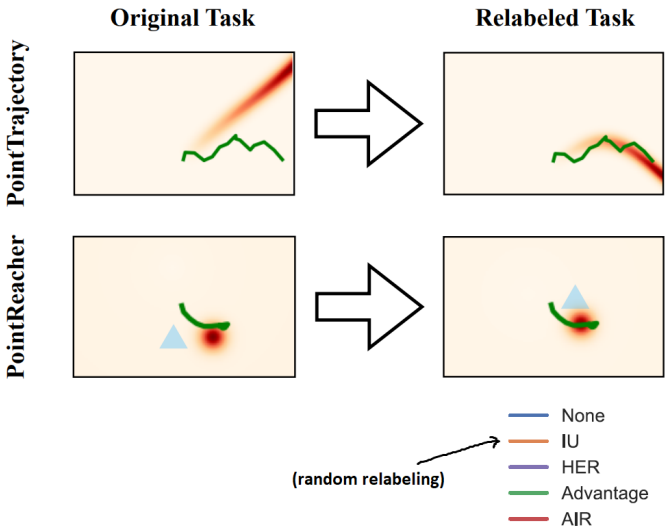
Advantage Relabeling

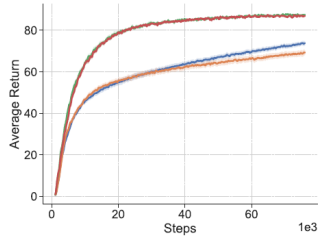


Advantage Relabeling

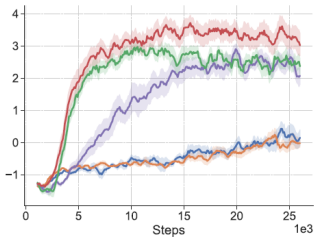


Generalized Hindsight: examples



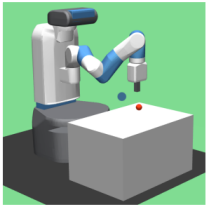


(a) PointTrajectory

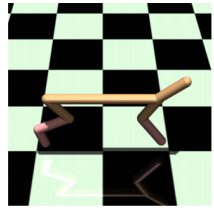


(b) PointReacher

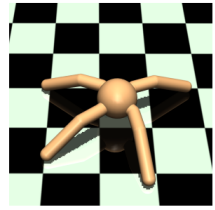
Generalized Hindsight: results



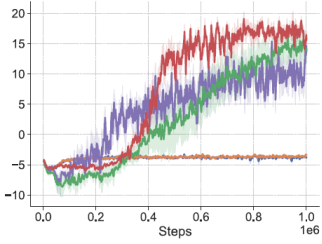
(c) Fetch



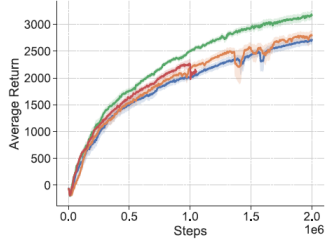
(d) HalfCheetahMultiObj



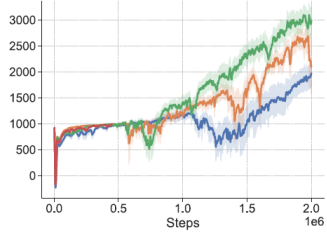
(e) AntDirection



(c) Fetch

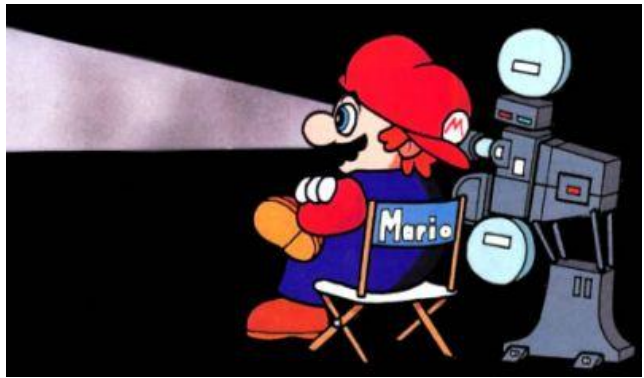


(d) HalfCheetahMultiObjective



(e) AntDirection

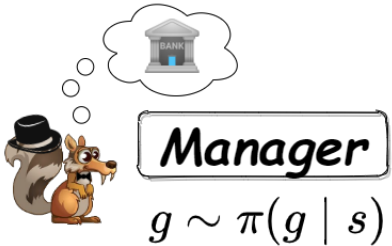
Generalized Hindsight: videos[4]



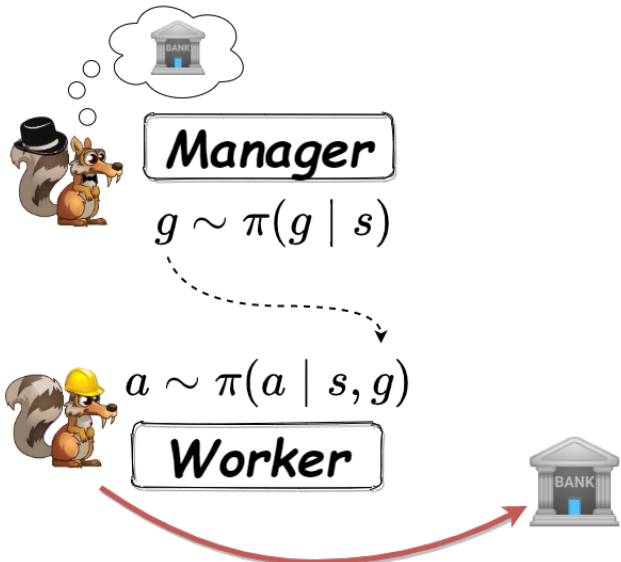
[4] https://www.youtube.com/watch?v=0E8T0Q5ZIYI

RL Dream: Hierarchical Reinforcement Learning

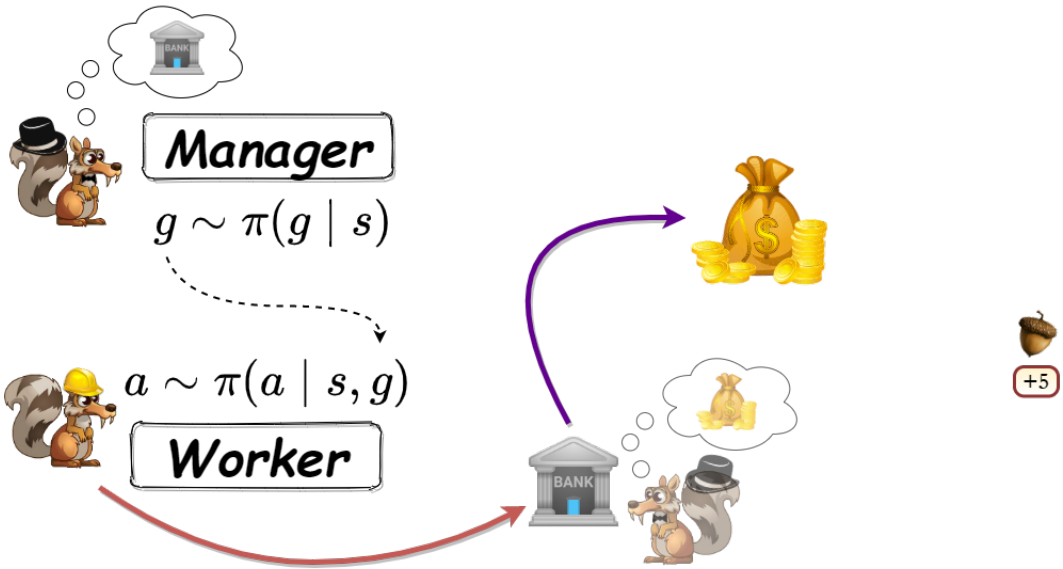
RL Dream: Hierarchical Reinforcement Learning



RL Dream: Hierarchical Reinforcement Learning



RL Dream: Hierarchical Reinforcement Learning



RL Dream: Hierarchical Reinforcement Learning