Learning Complementary Representations of the Past using Auxiliary Tasks in Partially Observable Reinforcement Learning

AAMAS 2020 — Auckland, New Zealand

Andrea Baisero Christopher Amato

{baisero.a,c.amato}@northeastern.edu

Northeastern University, Boston, USA

Northeastern University

AAMAS 2020

Overview

Setting:

- Partial observable reinforcement learning
- Single agent, model-free
- With memory requirements

Learning History Representations with Auxiliary Tasks:

- Train history representation $\phi(h)$ to help solve RL task
- Contributions:
  - Principles for good and efficient auxiliary tasks
  - Prediction-based auxiliary task which satisfies principles
  - “Complementary” architecture for training history representations with auxiliary tasks

Northeastern University

Background

POMDPs as History-MDPs

History-MDPs:

POMDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathrm{T}, \mathrm{O}, \mathrm{R} \rangle \Rightarrow$ History-MDP $\langle \mathcal{H}, \mathcal{A}, \mathrm{T}_\mathcal{H}, \mathrm{R}_\mathcal{H} \rangle$

- History states $\mathcal{H} \doteq (\mathcal{A} \times \mathcal{O})^*$
- History dynamics $\mathrm{T}_\mathcal{H} \colon \mathcal{H} \times \mathcal{A} \to \mathcal{H}$
- History rewards $\mathrm{R}_\mathcal{H} \colon \mathcal{H} \times \mathcal{A} \to \mathbb{R}$

Goal: Optimize parametric history policy $\pi_\mathcal{H} \colon \mathcal{H} \to \Delta\mathcal{A}$

Pros & Cons:

\+ Solve POMDPs using MDP methods

– History-states are hard
  - $|\mathcal{H}|$ exponential in horizon
  - Histories $h \in \mathcal{H}$ have different “sizes”
  - Same history never seen twice in one episode
  - Extremely hard to generalize

Northeastern University

Background

Internal State Representations

History Policy Decomposition: $\pi_{\mathcal{H}} \equiv \pi_{\mathcal{X}} \circ \phi$

- Internal-state set $\mathcal{X}$
- Internal-state representation $\phi\colon \mathcal{H} \to \mathcal{X}$
- Internal-state policy $\pi_{\mathcal{X}}\colon \mathcal{X} \to \Delta\mathcal{A}$

Common Internal State Representations

- Belief-state; $b(h) \in \Delta\mathcal{S}$
  - \+ Golden standard for generalization
  - – Requires known/learned model
- Reactive-m; concatenation of m latest interactions
  - \+ Great for short-term memorization
  - – Poor generalization for mid-/long-term
- Recurrent; recurrent neural network
  - \+ Potential for long-term memorization/generalization
  - – Hard to train to do so

Learning with Auxiliary Tasks

Motivation

Model-free RL trains ϕ and $\pi_\mathcal{X}$ using the RL objective

$$\phi^*, \pi_\mathcal{X}^* = \underset{\phi, \pi_\mathcal{X}}{\operatorname{argmax}} \mathbb{E}\left[G\right]$$

Problems:

- \+ Technically correct objective
- – Rewards are a weak training signal
- – Only implicit feedback on good representations ϕ
- – Sample experience contains more learning potential

Solution: Use an auxiliary task to train better ϕ

- \+ Fully exploit sample experience
- \+ Better representations, solve RL task better/faster

Northeastern University

Learning with Auxiliary Tasks

Principles

Goals:

1. Generalize like the true belief-state

$$b(h) = b(h') \Leftrightarrow \phi(h) = \phi(h')$$

2. Help $\pi_{\mathcal{X}}$ converge fast

Generalization Principles for Auxiliary Tasks

- Should be history-variant

$$h \not\approx h' \Rightarrow \phi(h) \not\approx \phi(h')$$

- Belief-state should be a sufficient statistic of history

$$b(h) \not\approx b(h') \Rightarrow \phi(h) \not\approx \phi(h')$$

Learning with Auxiliary Tasks

Principles

Goals:

1. Generalize like the true belief-state

$$b(h) = b(h') \Leftrightarrow \phi(h) = \phi(h')$$

2. Help $\pi_{\mathcal{X}}$ converge fast

Efficiency Guidelines for Auxiliary Tasks

- Should be "easier" than RL, e.g., self-supervised
  $\Rightarrow$ faster convergence
- Should be well-defined for every time-step
  $\Rightarrow$ data efficiency
- Should be stationary w.r.t. the agent
  $\Rightarrow$ sample efficiency (w/ experience replay)

Learning with Auxiliary Tasks

One-Step Predictive Task

One-Step Predictive Auxiliary Task (AUX)

Train ϕ and prediction model $p\colon \mathcal{X} \times \mathcal{A} \to \Delta\left(\mathcal{O} \times \mathcal{R}\right)$ to estimate observation-reward predictions

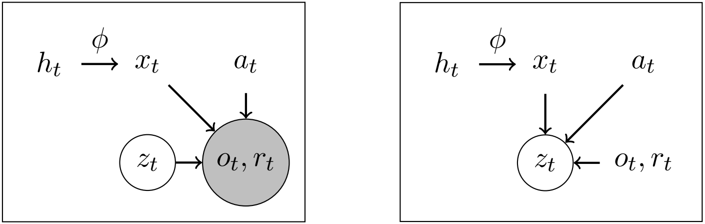
$$p(\phi(h), a) \mapsto \Pr(O, R \mid H = h, A = a) \,.$$

Advantages:

+ Satisfies principles & guidelines
  - History-variant
  - Belief-state as sufficient statistic
  - Self-supervised
  - Well-defined for every time-step
  - Stationary w.r.t agent
+ Based on observable data

Learning with Auxiliary Tasks

VAE for the One-Step Predictive Task



(a) Generative model $p(z, o, r; x, a)$ (b) Inference model $q(z; x, a, o, r)$

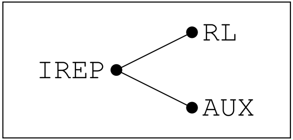
ELBO loss for observation-reward prediction

$$\mathcal{L}_{\text{ELBO}}(h, a, o, r) = \mathbb{E}_{z \sim q(z;x,a,o,r)} \left[\log \left(\frac{p(o, r; z, x, a)}{q(z; x, a, o, r)} \right) \right] \Bigg|_{x=\phi(h)}$$

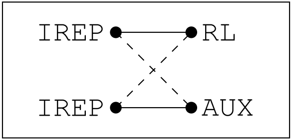
- Context variables: History h, action a
- Outcome variables: Observation o, reward r

Learning with Auxiliary Tasks

Learning Complementary Representations



(a) Representation w/ AUX (RAux)



(b) Complementary Representations w/ AUX (CRAux)

Left dots	Internal state representations
Right dots	Tasks
Solid edges	Representation used and trained on task
Dashed edge	Representation used but not trained on task

Northeastern University

Evaluation

Baselines

Baseline representations

TrueBelief	Belief-state representation of true model (as upper-bound on information-content)
React-{1,2,4}	Reactive representations w/ memory $\{1, 2, 4\}$
GRU	Recurrent representation

Proposed representations:

GRU-RAux	Recurrent representation trained w/ RAux
GRU-CRAux	Recurrent representation trained w/ CRAux

RL task solved using A2C + negative-entropy loss.

Northeastern University

Evaluation

Domains

Finite POMDPs w/ memory requirements:

Shopping-5 Localize and select the target item

Flexible task:

- Solvable with short-term memory
- Optimal solution requires mid-term memory

HeavenHell-3 Gather info and find the right exit.

Rigid task:

- Requires mid-term memory

RockSample-5-6 Find and collect good rocks

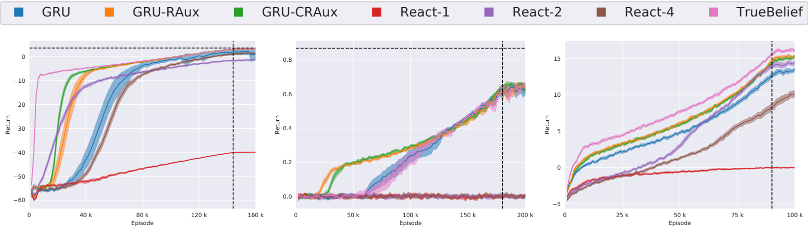
Larger and more stochastic task

- Solvable with short-term memory
- Optimal solution requires long-term memory

Northeastern University

Evaluation

Results



(a) Shopping-5. (b) Heavenhell-3. (c) Rocksample-5-6.

Figure: Training performance averaged over 40 independent runs, with shaded areas showing 2 standard errors of the mean.

Northeastern University

Conclusions

Summary

Conclusions

- RL task is insufficient to learn good representations $\phi(h)$
- Auxiliary tasks can train better representations $\phi(h)$
  $\Rightarrow$ help RL agent solve task better/faster

Contributions

- Principles for good and efficient auxiliary tasks
- Prediction-based auxiliary task which satisfies principles
- “Complementary” architecture for training history representations with auxiliary tasks

Northeastern University