Safety Aware Reinforcement Learning (SARL)

Santiago Miret *
santiago.miret@intel.com

Somdeb Majumdar *
somdeb.majumdar@intel.com

Carroll Wainwright †
carroll@partnershiponai.org

Introduction: As reinforcement learning (RL) agents become more and more prevalent in complex, real-world applications, the notion of safety becomes increasingly important. This has spurred a growth in research topics that incorporate safety considerations with RL algorithms [Amodei et al., 2016]. We focus specifically on minimizing side effects, where an agent's actions to perform a task in its environment may cause undesired, and sometimes irreversible, changes in the environment. Common measures of side effects are often task- and environment-specific with limited generalization to other settings. Previous work on formulating a generalized measure includes [Turner et al., 2019] on conservative utility preservation and [Krakovna et al., 2018] on relative reachability, which investigate abstract measures of side effects based on an analysis of changes in the state space. While these works motivate a better formulation of side effects, many studies have generally been limited to simple Gridworld [Leike et al., 2017] [Huang and Haskell, 2018] environments where the RL problem can often be solved in a tabular way and value function estimations are often not prohibitively demanding.

We consider more complex environments, generated by the SafeLife suite [Wainwright and Eckersley, 2020]. SafeLife creates complex environments of systems of cellular automata based on a set of rules from Conway's Game of Life [Gardner, 1970] that govern the interactions between, and the state (alive or dead) of, different cells. In addition to the basic rules, SafeLife enables the creation of complex, procedurally generated environments and patterns through special cells, such as *spawners* that can create new cells. Within the SafeLife suite, we focus on the *prune* task for the agent removes cells from the board, and the *append* where the agent is building patterns.

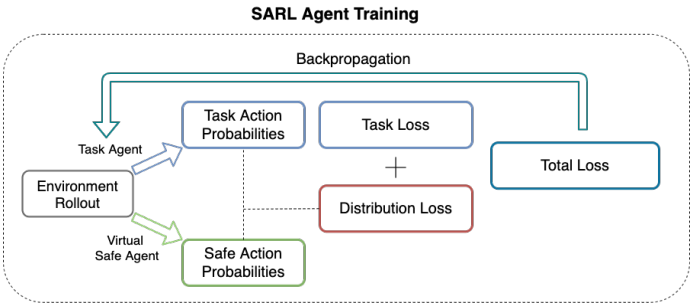


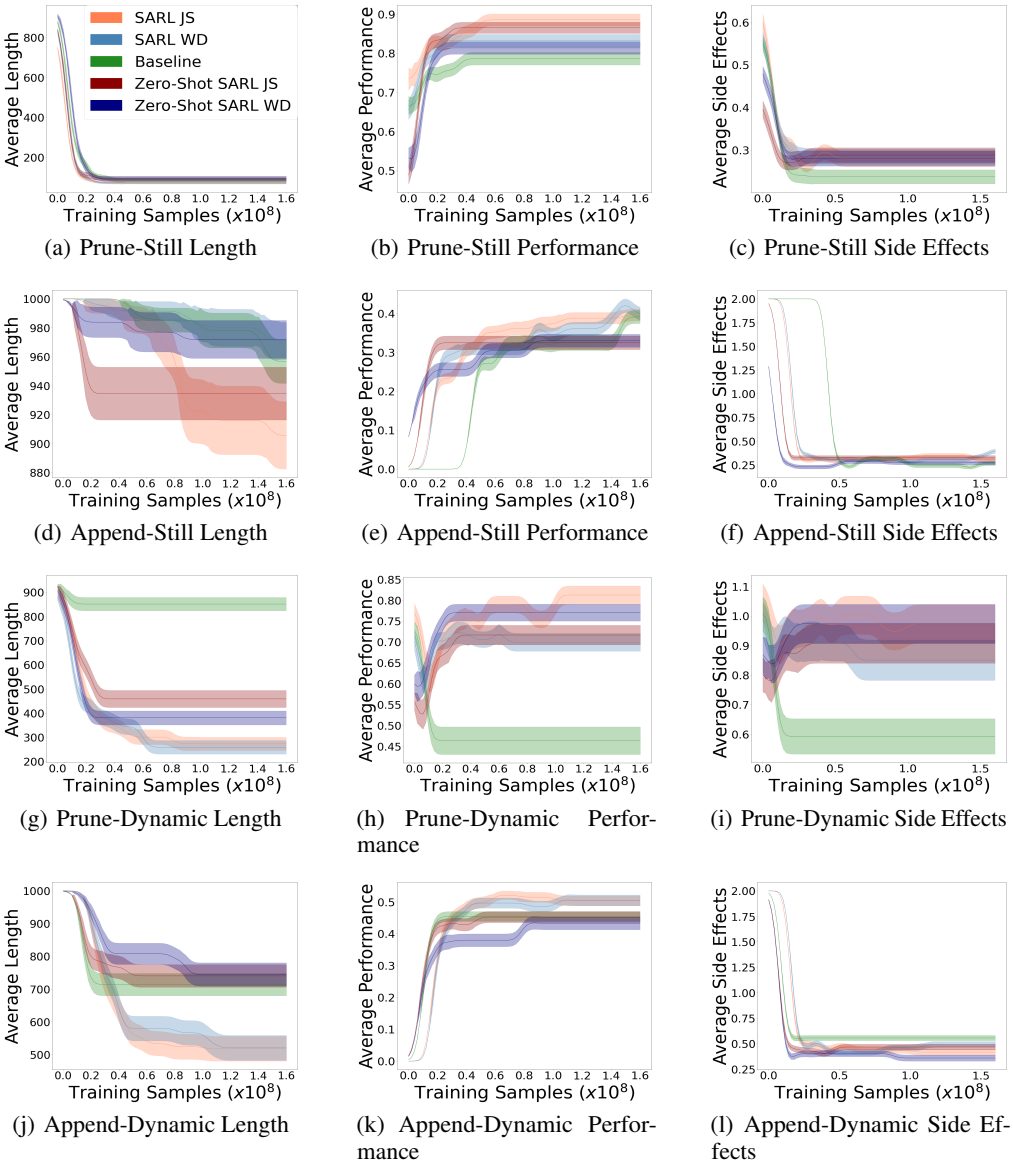
Figure 1: Co-Training Framework for SARL

Method: We present Safety Aware RL (SARL) a novel architecture that trains an RL agent using a primary task objective and simultaneously modulates its behavior via a secondary RL agent. The safety agent inside SARL computes a generalized measure of safety by representing it as a value function approximation of an environment-specific side effect metric. The primary agent's objective is then regularized by a loss-formulation that is a function of the distance between its own value function estimates and that of the safe actor. We use the Jensen-Shannon Distance and the dual formulation of the Wasserstein Distance described in Pacchiano et al. [2019] as distance metrics between action probabilities based on the value function estimations. We also hypothesize that this abstracted notion of

*Intel Labs
†Partnership on AI

LatinX in AI Workshop at Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

safety can generalize to multiple tasks in the same environment. As a baseline, we compare with [Wainwright and Eckersley, 2020] where the primary agent's objective was regularized directly using a task-specific side-effect defined directly in the raw observation space. Our goal is to show that the performance of the abstracted formulation is competitive with the baseline for a given task and can zero-shot generalize to new tasks with no further training of the safe actor. More formally, the general objective of the task agent A_θ can be expressed as: $\mathcal{F}_\mathcal{A}(\theta) = \mathcal{L}_\theta + \beta * \mathcal{L}_{\text{dist}}(\mathbb{P}_{\pi_\theta}, \mathbb{P}_{\pi_\psi})$, where β is a regularization hyperparameter, $\mathbb{P}_{\pi_\theta}$ represents the probability of taking a given action given by A_θ, and $\mathbb{P}_{\pi_\psi}$ represents the probability distribution of taking a given action according to Z_ψ.



(a) Prune-Still Length (b) Prune-Still Performance (c) Prune-Still Side Effects

(d) Append-Still Length (e) Append-Still Performance (f) Append-Still Side Effects

(g) Prune-Dynamic Length (h) Prune-Dynamic Performance (i) Prune-Dynamic Side Effects

(j) Append-Dynamic Length (k) Append-Dynamic Performance (l) Append-Dynamic Side Effects

Figure 2: Length Champion for SafeLife Suite of *1. prune-still (a-c), 2. append-still (d-f), 3. prune-dynamic (g-i), 4. append-dynamic (j-l)* tasks evaluated for 100 testing environments every 100,000 steps on *Episode Length (left column)* where shorter is better, *Performance Ratio (middle column)* where higher is better, and *Episodic Side Effect (right column)* where lower is better

The results of the experiments shown in Figure 2 demonstrate that a virtual safety agent trained on one task in the SARL framework can generalize zero-shot to other tasks and environment settings, while maintaining competitive task and side effect scores compared to the baseline method. This allows us to abstract the notion of safety away from the environment specific side effect metric

References

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Manè. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

M. Gardner. The fantastic combinations of jhon conway's new solitaire game'life. *Sc. Am.*, 223: 20–123, 1970.

W. Huang and W. B. Haskell. Stochastic Approximation for Risk-aware Markov Decision Processes. pages 1–34, 2018. URL `http://arxiv.org/abs/1805.04238`.

V. Krakovna, L. Orseau, R. Kumar, M. Martic, and S. Legg. Penalizing side effects using stepwise relative reachability. *arXiv preprint arXiv:1806.01186*, 2018.

J. Leike, M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, and S. Legg. Ai safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

A. Pacchiano, J. Parker-Holder, Y. Tang, A. Choromanska, K. Choromanski, and M. I. Jordan. Learning to Score Behaviors for Guided Policy Optimization. 2019. URL `http://arxiv.org/abs/1906.04349`.

A. M. Turner, D. Hadfield-Menell, and P. Tadepalli. Conservative agency. *arXiv preprint arXiv:1902.09725*, 2019.

C. L. Wainwright and P. Eckersley. SafeLife 1.0: Exploring side effects in complex environments. *CEUR Workshop Proceedings*, 2560:117–127, 2020. ISSN 16130073.