Safety Aware Reinforcement Learning (SARL)

Santiago Miret (Intel Labs), Somdeb Majumdar (Intel Labs), Carroll Wainwright (PAI)

Safety in Reinforcement Learning:

As we aim to deploy RL agents in the real world, safety considerations will become a more and more important. As such, researchers have created environments for practitioners to experiments with safety considerations for RL agents, such as DeepMind's SafetyGrids and OpenAI's Safety Gym. Among the plethora of safety challenges within RL[1], we focus on side effects.

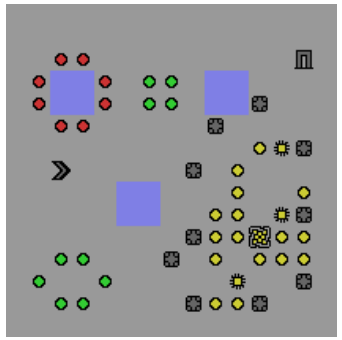
SafeLife:[2]

SafeLife is a suite of environments built upon the three of Conway's Game of Life:

1. Every cell is either alive or dead.
2. Live cells die unless they have 2 or 3 neighbors.
3. Dead cells stay dead unless they have exactly 3 neighbors

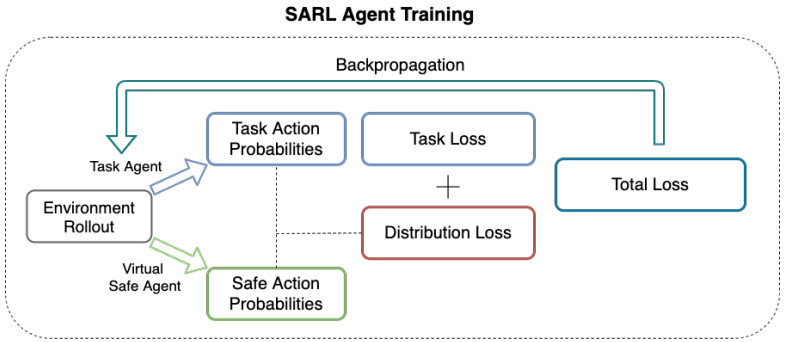
Based on those rules, SafeLife incorporates intricate dynamic changes into its environment, enabling an agent to interact with an environment where its actions can have anticipated consequences and measurable side effects.

Sample environment in the SafeLife suite: The agent (arrow symbol) performs either prune (removing red cells) or append task (filling in blue cells) before exiting through a level exit (door symbol)



SARL:

SARL operates by regularizing the loss of the RL agent with a probabilistic distance metric. The probabilistic distance is computed between a distribution describing task efficient behavior ($P_{\pi_\theta}^{task}$) stemming from a task policy π_θ and safety efficient behavior ($P_{\pi_\psi}^{safety}$) stemming from a safe policy π_ψ. In our paper we apply the Jensen-Shannon Distance and a dual formulation of the Wasserstein Distance[3] to compute the difference between the two distributions.



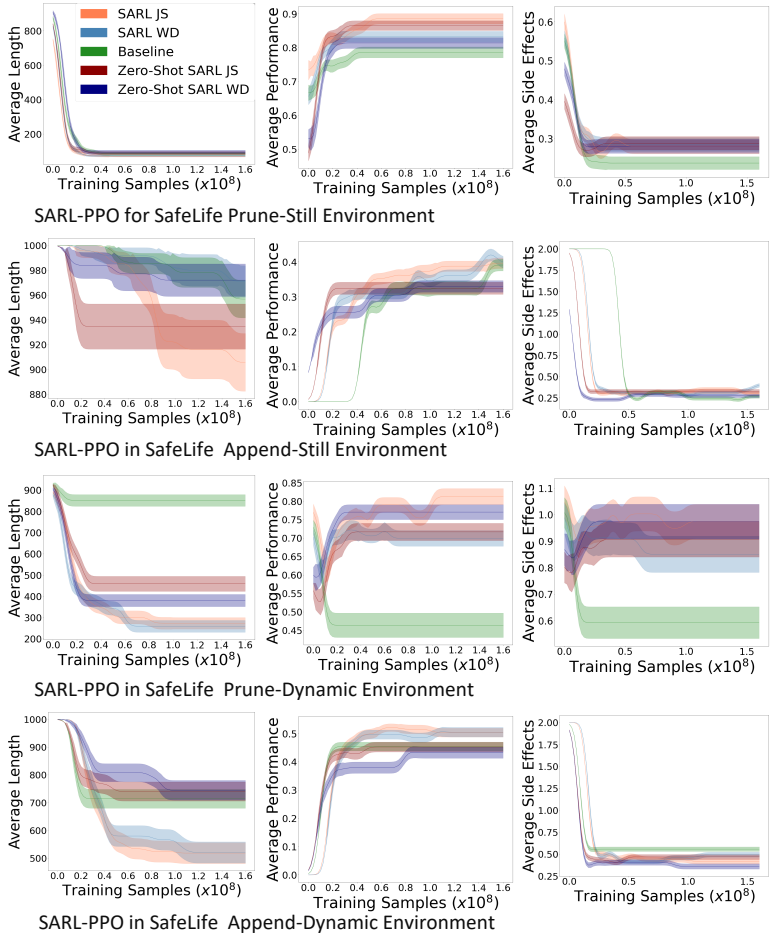
Experiments:

We conduct experiments on four environments in the SafeLife Suite: prune-still, append-still, prune-dynamic, append-dynamic. We modify the PPO[4] algorithm to integrate into the SARL framework and conduct two types of training of SARL-PPO:

- Both the task agent and the safety agent are trained from scratch
- The safety is trained on another environment and applied without changes (zero-shot) to the environment the task agent is training on.

Results:

We show results for our experiments and observe that the safety agent in SARL-PPO can generalize zero-shot to new environments. We show episode length (left), performance defined by $\frac{reward\_achieved}{reward\_possible}$ (middle), and side effects (right)



References:
1. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Manè. Concrete problems in aisafety.arXiv preprint arXiv:1606.06565, 2016.
2. C. L. Wainwright and P. Eckersley. SafeLife 1.0: Exploring side effects in complex environments.CEUR Workshop Proceedings, 2560:117–127, 2020. ISSN 16130073.
3. A. Pacchiano, J. Parker-Holder, Y. Tang, A. Choromanska, K. Choromanski, and M. I. Jordan.Learning to Score Behaviors for Guided Policy Optimization. 2019. URLhttp://arxiv.org/abs/1906.04349.
4. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimizationalgorithms.arXiv preprint arXiv:1707.06347, 2017.