Creative part: approach

CSAI: Contemporary AI Topics

December 5, 2019

Robbe Louage & Elias Storme

We want to demonstrate the performance of both approaches to safety. This requires a level playing field, so results can be compared. This leveling of the playing field is a recurring problem in AI research. To solve this problem, shared datasets have emerged and been adopted by researchers. One well known example of this is ImageNet, a database of images to do image processing on.

1 DeepMind Gridworlds

For safety in reinforcement learning, one such "dataset" being proposed is Gridworlds by Google DeepMind [3]. There are a total of 8 gridworlds. Each gridworld represents a different environment that can be used to test one aspect of AI safety. The environments include the following safety checks for a learning agent:

- Safe interruptibility
- Avoiding side effects
- Absent supervisor
- Reward gaming
- Self-modification
- Distributional shift
- Robustness to adversaries
- Safe exploration

In the following section we will discuss how we will evaluate both systems on the previous criteria.

2 Approach

In order to familiarize ourselves more with the particulars of the papers [1] and [2], we will first explore their given implementations. Both papers include their respective approach implemented on a variety of games.

To evaluate performance of both ways to implement safety, we of course need a learner as baseline. We will look at the learners used in both papers, as well as the learners being used in research today. From those we will select one which is suitable to implement both approaches on.

Having obtained a learner, we will of course implement both approaches. Doing this we will evaluate and report on the ease of use of each approach. Then we will train them on the Gridworlds environment. The first criterion for evaluation is convergence. Both safety approaches are guaranteed to converge on MDPs on which the learner converges without safety. Thus, we will be looking at the speed of convergence. Secondly, we will be evaluating the checks outlined above.

References

[1] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[2] Giuseppe De Giacomo, Luca Iocchi, Marco Favorito, and Fabio Patrizi. Foundations for restraining bolts: Reinforcement learning with ltlf/ldlf restraining specifications. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, pages 128–136, 2019.

[3] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. Ai safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.