Master Research Proposal Nr:

Safe Reinforcement learning using Successor representations

Reinforcement Learning (RL) has achieved remarkable success in various domains, from game playing to robotics. However, in safety-critical applications such as autonomous driving or medical treatment, traditional RL algorithms may lead to unsafe behavior. Safe RL aims to address this challenge by ensuring that agents operate within predefined safety constraints while maximizing their objectives.

Successor Representations typically split the problem of learning the value of any individual state into two components. First, we have a reward prediction function which simply tries to predict the immediate utility of every state. Second, we have a successor map that predicts the future state occupancy rate when starting from the current state [1]. Most Constrained Reinforcement learning algorithms, such as CPO [2] or Chow's CVAR method [3], rely on estimating state visitation probabilities either explicitly or implicitly. The reason for this is because to compute whether a policy violates a constraint, they need to know the likelihood a violating state is reachable by the current policy.

Successor representations (SR) offer a promising approach to model the dynamics of the environment by estimating the probability distribution of reaching a state at any point in the future given the current state and action [1,4,5,6]. Safe RL requires methods to ensure that agents operate within safe regions of the state space. This is typically achieved by defining a safety function that assigns a safety score to each state, and agents must avoid states with negative safety scores. The successor function $M(s,s')$ can be used to determine whether a previously discovered unsafe state is reachable via the new policy. Current methods mostly rely on either Monte-Carlo estimating state visitation probabilities [2], or computing worst-case bounds based on those estimates [3]. As is usual with Monte Carle methods, such estimates are characterized by a high variance, which results in high sampling complexity and/or in and under-estimation of the true risk.

The goal of this thesis is to instead of using Monte-Carlo to estimate the successor distribution, to learn such a distribution to yield a lower-variance estimator for policy constraint violations. Since SR decomposes the discounted constraint function and the value function into a visitation rate and a state-wise function [1,4,5,6] one can re-use the visitation rate for both value and constraint estimation. This is because the Value function V(s) and cumulative constraint function $C_i(S)$ can be represented as an integral $V(S) = \int_S M(s,s')R(s')ds'$ or $C_i(S) = \int_S M(s,s')c_i(s')ds'$, respectively, where the integral can be approximated through a replay buffer of previously seen states. This reduces the usage of the value function and (multiple) constraints to a single re-weighting of the existing replay buffer by $M(s,s')$. This decomposition also means that constraints/values can be added or altered online (see, for instance, [5,6]).

The resulting lower-variance algorithm should be tested against existing CMDP algorithms (e.g.,CPO[2], Chow's CVAR[3] method, Lagrangian penalty) on representative environments in safety gymnasium (e.g., Push, Circle, Goal).

The primary objective of this research is to develop novel algorithms that combine successor representations with safety-aware reinforcement learning techniques. Specifically, the time plan is as follows:

- Literature review (1 month)
- Algorithm implementation and tuning (2 months)
- Benchmarking on safety gym (2 month)
- Writing the thesis (1 month)

The thesis must contain a detailed description of all developed and used algorithms as well as a profound result evaluation and discussion. The implemented code must be documented and provided.

Advisors: Alexander Mattick, Dr. Christopher Mutschler, Prof. Dr. Björn Eskofier
Student: Michael Girstl
Start—End: 1.11.2024 – 30.04.2025

References

[1] *Kulkarni, T.D., Saeedi, A., Gautam, S., & Gershman, S.J. (2016). Deep Successor Reinforcement Learning. ArXiv, abs/1606.02396.*

[2] Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained Policy Optimization. ArXiv, abs/1705.10528.

[3] Chow, Y., Ghavamzadeh, M., Janson, L., & Pavone, M. (2015). Risk-Constrained Reinforcement Learning with Percentile Risk Criteria. ArXiv, abs/1512.01629.

[4] Gershman, S. J. (2018). The Successor Representation: Its Computational Logic and Neural Substrates. The Journal of neuroscience : the official journal of the Society for Neuroscience 38, 7193–7200. doi:10.1523/JNEUROSCI.0151-18.2018.

[5] Ma, C., Wen, J., and Bengio, Y. (2018). Universal Successor Representations for Transfer Reinforcement Learning. arXiv:1804.03758 [cs, stat]. Available at: https://arxiv.org/abs/1804.03758.

[6] Zhang, J., Springenberg, J. T., Boedecker, J., and Burgard, W. (2016). Deep Reinforcement Learning with Successor Features for Navigation across Similar Environments. arXiv:1612.05533 [cs]. Available at: https://arxiv.org/abs/1612.05533.