A Distributional Perspective on Reinforcement Learning

Manan Tomar : ED14B023

Abstract

This work (Bellemare et al., 2017) proposes a distributional perspective to reinforcement learning by introducing learning over the random return instead of its expected value and argues that doing so allows for preserving mutimodality, resulting in more stable learning.

1. Approach

Traditionally, Reinforcement Learning focuses on estimating the expected return, also termed as the $Q(s, a)$ or $V(s)$ value (Sutton et al., 1998). This work explores estimating the overall distribution of the random return variable, instead of just the expected value. This is called as the value distribution. Similar to the bellman equation, the authors introduce an analogous equation for the value distribution $Z(s, a)$ as follows.

$$Z(s, a) = R(s, a) + \gamma Z(s^{'}, a^{'}) \tag{1}$$

,where the equality is between two distributions and not scalars.

The paper then moves on to proving that the analogous bellman operators $\mathcal{T}$, $\mathcal{T}^{\pi}$ are contraction mappings under the Wasserstein metric. The authors then propose a practical algorithm for implementing the above. Two main contributions here are:

- The value distribution is considered as a discrete distribution parameterized by N supports, where the probability of each category bin is defined by the softmax over the output of a θ parameterized model $\theta : \mathcal{X} \times \mathcal{A} \rightarrow R^n$. The supports are bounded between a user selected $V_{\text{MIN}}, V_{\text{MAX}}$.
- The cross entropy loss between the two distributions present in the bellman equation (represented by the LHS and RHS) is used to minimize the temporal difference error.

During each update, the sampled transitions from the environment are used to extimate the target value distribution $Z^{'}$ and use it to update Z using the cross entropy loss.

2. Experiments

The authors show results on the Atari domain and compare with similar works in DQN (Mnih et al., 2015), Double DQN, Prioritized Experience Replay and Duel DQN. The number of bins used to represent the value distribution is fixed to 51 for all experiments. It is observed that C-51 performs almost as good or better than the above baselines. The performance gap is especially significant in sparse reward games.

According to the authors, the major benefits provided by this method are as follows.

- Using a value distribution allows for representing multimodal distributions, in which case using the expected value can be sub optimal.
- This also allows for risk averse behavior by choosing an action which has less variance in the case two actions have similar expected values.
- The clipping on supports of the value distribution provide a good way to incorporate domain knowledge in the learning problem.

Finally, although the algorithm is provided as one which can be practically implemented, the choice of using 51 supports to model the value distribution is completely empirical and the performance is affected significantly for different values. Moreover, the authors only use a discrete distribution while the proposed algorithm can potentially be more powerful if used with more complex distributions.

References

Bellemare, Marc G, Dabney, Will, and Munos, Rémi. A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*, 2017.

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A, Veness, Joel, Bellemare, Marc G, Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K, Ostrovski, Georg, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Sutton, Richard S, Barto, Andrew G, Bach, Francis, et al. *Reinforcement learning: An introduction*. MIT press, 1998.