Review-37

Transfer in Deep Reinforcement Learning Using Successor Features and Generalised Policy Improvement

Successor Features (SFs) in conjunction with Generalized Policy Improvement (GPI) have enabled successful transfer of skills in Reinforcement Learning (RL) agents between tasks. However, the SF&GPI framework assumes rewards as linear combinations of fixed set of features. The work presented in the paper rerlaxes this constraint and generalizes SF&GPI to any set of tasks that only differ in the reward function. One can use reward functions themselves as features for future tasks, thus reducing the need to specify features beforehand. The proposed scheme learns online and demonstrates instantaneous transfer performance on challenging 3D environments in comparison to baseline methods.

The work generalizes the framework of SF&GPI by inducing a set of MDPs using successor features which are linearly independent of each other. This allows rewards to be expressed as linear combinations of successor features and weight vectors serving as approximations under some well-defined criterion. The framework further presents that rewards can themselves be used as features since successor feature vectors are a linear representation of rewards. Such a formulation allows for a more convenient approximation which can be easily learned in the more general case. The generalized SF&GPI method is implemented using ϵ-greedy Q-learning by treating Q-values as inner products of successor feature sets and weight vectors. Incorporating rewards as features allows for simpler approximations in this algorithm by directly updating the policy with respect to reward approximations.

Proposed SF&GPI framework demonstrates improved, in some cases instantaneous, transfer performance on test tasks when executed on challenging 3D environments. Another notable finding is that the agent is able to perform well on tasks with negative rewards even though it was trained only on tasks with positive rewards. This indicates that the transfer policy is successfully combining policies corresponding to base tasks. The work further compares the generalized framework with prior method of linearly-dependent base tasks. The generalized framework, having an upper bound on Q-value errors, presents sub-optimal performance in comparison to linear base tasks. Furthermore, the scheme extends utilization of rewards as features based on significant assumptions which may not hold true in practical scnearios. For instance, transferring of features assumes that base tasks as linearly independent of each other which may not be true in the case of temporally-extended environments. Lastly, scheme takes into account various sources of error but does not throw light on approximation bias of Q-values which comprise of feature and weight vector estimates. The question of whether this induces inherent biases in estimates remains as open problem.