Review-34

The Value-Improvement Path Towards Better Representations for Reinforcement Learning

Unlike supervised learning, value-based Reinforcement Learning (RL) pertains to a sequence of non-stationary problems of finding the optimal value function. An alternate viewpoint consists of formulating the setup as a single holistic prediction problem. The RL agent optimizes its policies towards an optimal policy by following an associated sequence of value functions collectively forming the *value improvement path*. The proposed approximation of value improvement path leads to representations which accurately express future policy improvements. The hypothesis leverages auxilary tasks to demonstrate the suitability of learning informative representations.

In contrast to contemporary RL, the proposed hypoethsis explicitly characterizes the observed sequence of value functions as value improvement path. Representations which accurately approximate past value improvements may also do so for future functions of the given path. This insight may be illustrated by collecting policy iteration's value improvement paths in a graph with the root node denoting the optimal value function. Policy improvement fulfills the objective of reaching the root node starting from a given leaf node denoting an arbitrary initialized value. The task of seeking a suitable value function comprises of informative representations which are in turn used to approximate future values. Thus, generalization of agent's values is directly posed as the learning of representations along the value improvement path. In order to support the theoretical claims, the work proposes learning of auxilary tasks such as cumulative value functions, cumulative policies and past policies in the value improvement path. When compared to the standard setting of learning only values, these form the basis of empirical analysis.

Suitability of past policies and past mixtures as auxilary tasks for optimizing value improvement path is validated upon observing increased generalization in comparison to other baseline methods. The proposed auxilary tasks denote lower errors and increased normalized scores on held out transitions in the case of Atari benchmark. Performance trends denote the method's ability to generalize to future value functions and its implication of the error being predictive of long-term performance. However, the value improvement path does not take into account various other auxilary tasks which are closely related with the value function and may yield a strong connection with the learning of informative representations. For instance, the most popular of auxilary objectives make use of maximum entropy and intrinsic exploration in order for the agent to approximate suitable functions in the value space. The value improvement path does not throw light in this regard and its compatibility to moer complex objectives remains an open question.

The work presents a theoretical alternative towards value-based learning of agents in RL by formulating the problem as a holistic prediction approach. The avenue presents two new directions for future work with the first being its scalability to complex objectives such as intrinsic motivation. Secondly, the approach may be furthr coupled with policy search in the form of policy gradient methods which may be realized as a prediction problem requiring traversal in the policy space.

Value-based RL consists of prediction and control problems with the former estimating long-term values and the latter optimizing them in light of current policy. The proposed approach reformulates value-based RL as a holistic prediction problem with the agent traversing the value space in search of the optimal value function. The value improvement path thus produced provides an effective mechanism for learning informative representations which aid generalization upon aligning future predictions along the path. Empirical analysis of the proposed method on Atari benchmark demonstrates its improved generalization and efficacy when compared to baseline methods.