Contribution ID: **33** Type: **Poster**

Testing and Improving RL Policies via Rule Learning

Tuesday, 31 March 2026 12:00 (2 hours)

Using domain knowledge to improve deep RL policies is a current challenge. LEGIBLE mines rules from an RL policy, constituting a partially symbolic representation. These rules describe which decisions the RL policy makes and which it avoids making. It then generalizes the mined rules using domain knowledge. Finally, it evaluates generalized rules to determine which generalizations improve performance when enforced. These improvements show weaknesses in the policy, where it has not learned the general rules and thus can be improved by rule guidance. We show the efficacy of our approach by demonstrating that it effectively finds weaknesses, accompanied by explanations of these weaknesses in several RL environments.
Closing the loop from neural to symbolic and back to neural representation, we show how to integrate symbolic (rule-based) knowledge into neural RL policies by leveraging RL from demonstrations with OFTEN-DeepRL.

Student

Yes

Primary author: LOPEZ-MIGUEL, Ignacio D.

Presenter: LOPEZ-MIGUEL, Ignacio D.

Session Classification: Poster session