Logic-Based Sequential Decision-Making

Daoming Lyu,[1] **Fangkai Yang,**[2] **Bo Liu,**[1] **Daesub Yoon**[3]
[1]Auburn University, Auburn, AL, USA
[2]Maana Inc., Bellevue, WA, USA
[3]ETRI, Daejeon, South Korea
{daoming.lyu, boliu}@auburn.edu, wolfgang.yang@gmail.com, eyetracker@etri.re.kr

Introduction

Deep reinforcement learning (DRL) has achieved great success on sequential decision-making problems involving high-dimensional sensory inputs such as Atari games (Mnih et al. 2015). The input states of Atari games are usually raw pixel images, and a deep neural network is used to approximate Q-values, i.e., "Deep Q-Network" (DQN). This approach can learn fine granular policies that surpass human experts but is criticized for the lack of data efficiency and explainability. DRL algorithms usually require several millions of samples but still cannot learn long-horizon sequential actions for problems with sparse feedback and delayed rewards, such as Montezuma's Revenge (Mnih et al. 2015). The learning behavior based on the black-box neural network is nontransparent and hard to explain and understand. In real applications of decision-making, however, it is instrumental to make the system behavior explainable to gain the trust from the user and provide insights for their decision-making process (Gilpin et al. 2018) with reasonable less data samples.

In this paper, we introduce the logic-based approach into DRL and propose a framework of Symbolic Deep Reinforcement Learning (SDRL) by utilizing *Symbolic Planning* (SP) (Cimatti, Pistore, and Traverso 2008), which can improve the data-efficiency and explainability of DRL. In addition, we propose the *intrinsic goal*, a measurement of plan quality based on an internal utility function, to enable reward-driven planning.

SDRL Framework

SDRL framework can handle both high-dimensional sensory inputs and symbolic planning, and features a *planner – controller – meta-controller* architecture, as shown in Figure 1.

With a symbolic representation given by the human expert, a symbolic planner generates high-level plans, i.e., a sequence of subtasks, to meet its *intrinsic goal*. An intrinsic goal is a measurement on plan quality, which approximates how much cumulative reward the plan may achieve. We assume a pre-trained mapping function can associate each sensory input with a symbolic state, i.e., performing symbol grounding, so that a set of options on the problem MDP can be induced based on symbolic states and the mapping

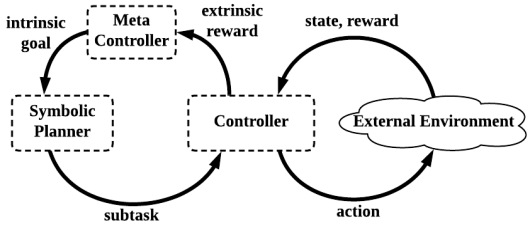


Figure 1: Architecture illustration

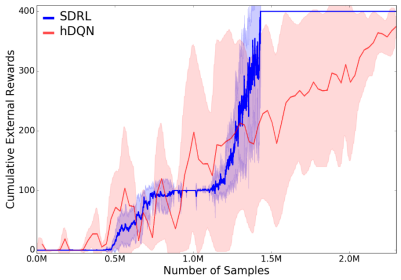
function. We extend the reward structure of core MDP by introducing *intrinsic reward* and *extrinsic reward* to facilitate two levels of learning tasks. The sub-policies for the action level are learned using DRL algorithms based on intrinsic reward, with pseudo-rewards to encourage the agent to learn skills to achieve each subtask. As DRL continues, a metric is used to evaluate the competence of learned sub-policies, such as the success ratio over a number of episodes, from which extrinsic rewards is derived. When the sub-policy is learned and reliably achieves the subtask, the extrinsic reward is equivalent to the environmental reward. Using extrinsic rewards, meta-controller performs R-learning that reflects the long-term average reward and gains the reward of selecting each subtask. The learned values are returned to the symbolic planner and are used to measure plan quality and propose new intrinsic goals for the planner to improve the plan, by either exploring new subtasks or by sequencing learned subtasks that supposedly can achieve higher rewards in the next iteration.

In this process, the components of planner, controllers, and meta-controller cross-fertilize each other and eventually converge to an optimal symbolic plan along with the learned subtasks. While our framework is generic enough so that various planning and DRL techniques can be used, we instantiate our framework using action language $\mathcal{BC}$ for planning and R-learning for meta-controller learning.

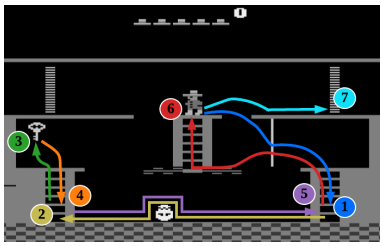
Experiment

The proposed approach is evaluated on Taxi domain and Montezuma's Revenge, demonstrating improved explainability through explicitly encoding planning knowledge and learning into human-readable subtasks, and also improved data-

efficiency through automatic selecting and learning control policies of modular subtasks. Due to the space, some experimental results of Montezuma's Revenge are shown in Figure 2, while the detailed ones of both Taxi domain and Montezuma's Revenge can be found in the supplement.



(a) Learning Curve (Montezuma's Revenge)



(b) Final Solution (Montezuma's Revenge)

Figure 2: Experimental Results

No.	subtask	policy learned	in optimal plan
1	MP to LRL, no key	✓	✓
2	LRL to LLL, no key	✓	✓
3	LLL to key, no key	✓	✓
4	key to LLL, with key	✓	✓
5	LLL to LRL, with or without key	✓	✓
6	LRL to MP, with or without key	✓	✓
7	MP to RD, with key	✓	✓
8	LRL to LS, with or without key	✓	
9	LS to key, with or without key	✓	
10	MP to RD, no key	✓	
11	LRL to key, with or without key		
12	key to LRL, with key		
13	LRL to RD, with key		

Table 1: Subtasks for Montezuma's Revenge

We formulated domain knowledge of Montezuma's Revenge in action language $\mathcal{BC}$ based on 6 pre-defined locations: middle platform (`mp`), right door (`rd`), left of rotating skull (`ls`), lower left ladder (`lll`), lower right ladder (`lrl`), and key (`key`). All the subtasks are shown in Table 1 as well. In Figure 2a, our approach (SDRL) is compared with hierarchical DQN (hDQN) (Kulkarni et al. 2016), which is set as the baseline. The learning curve of SDRL shows that the agent first discovered the plan of collecting key after 0.5M samples by sequencing subtasks 1–3. Intrinsically motivated planning encourages exploring untried subtasks, and by learning more subtasks to move to other locations, the agent finally converges to the maximal cumulative external reward of 400 around 1.5M samples by sequencing subtasks 1–7 (Fig. 2b). By comparison, hDQN cannot reliably achieve the score of 400 around 2.5M samples. The variance of SDRL is smaller than that of hDQN, partially due to the fact that our definition of subtask is easier to learn than the one defined in hDQN, leading to more robust and stable learning. During the experiment, subtasks 1–10 are successfully learned by DQNs, with 7 of them being selected in the final solution. It should be noted that the order that subtasks are learned does not depend on the order they appear in the final optimal plan, and the subtasks proposed for learning by symbolic planner is activated only when the starting state is satisfied, and once learned, can be easily sequenced and reused in other plans.

Conclusions

In this paper, we propose SDRL framework by integrating symbolic planning with deep reinforcement learning for decision making. The task-level explainability is enabled by relating symbolic actions to options.This framework takes charge of subtask scheduling, data-driven subtask learning, and subtask evaluation alternatively, and makes the final solution converge to an optimal symbolic plan along with the learned subtasks, bringing together the advantages of long-term planning capability with symbolic knowledge and end-to-end reinforcement learning directly from a high-dimensional sensory input. Experimental results validate the explainability of subtasks, along with improved data efficiency compared with state-of-the-art approaches.

Acknowledgments

Daoming Lyu, Bo Liu, and Daesub Yoon were partially supported by a grant (18TLRP-B131486-02) from Transportation and Logistics R&D Program funded by Ministry of Land, Infrastructure and Transport of Korean government. We also thank the donation of GPU card from NVIDIA Inc.

References

Cimatti, A.; Pistore, M.; and Traverso, P. 2008. Automated planning. In van Harmelen, F.; Lifschitz, V.; and Porter, B., eds., *Handbook of Knowledge Representation*. Elsevier.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*.

Kulkarni, T. D.; Narasimhan, K.; Saeedi, A.; and Tenenbaum, J. 2016. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in Neural Information Processing Systems*, 3675–3683.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.