Learning to Play Text-based Adventure Games with Maximum Entropy Reinforcement Learning

Weichen Li[1], Rati Devidze[2], Sophie Fellenz[1]

[1] University of Kaiserslautern-Landau, Germany [2] Max Planck Institute for Software Systems, Germany

Abstract

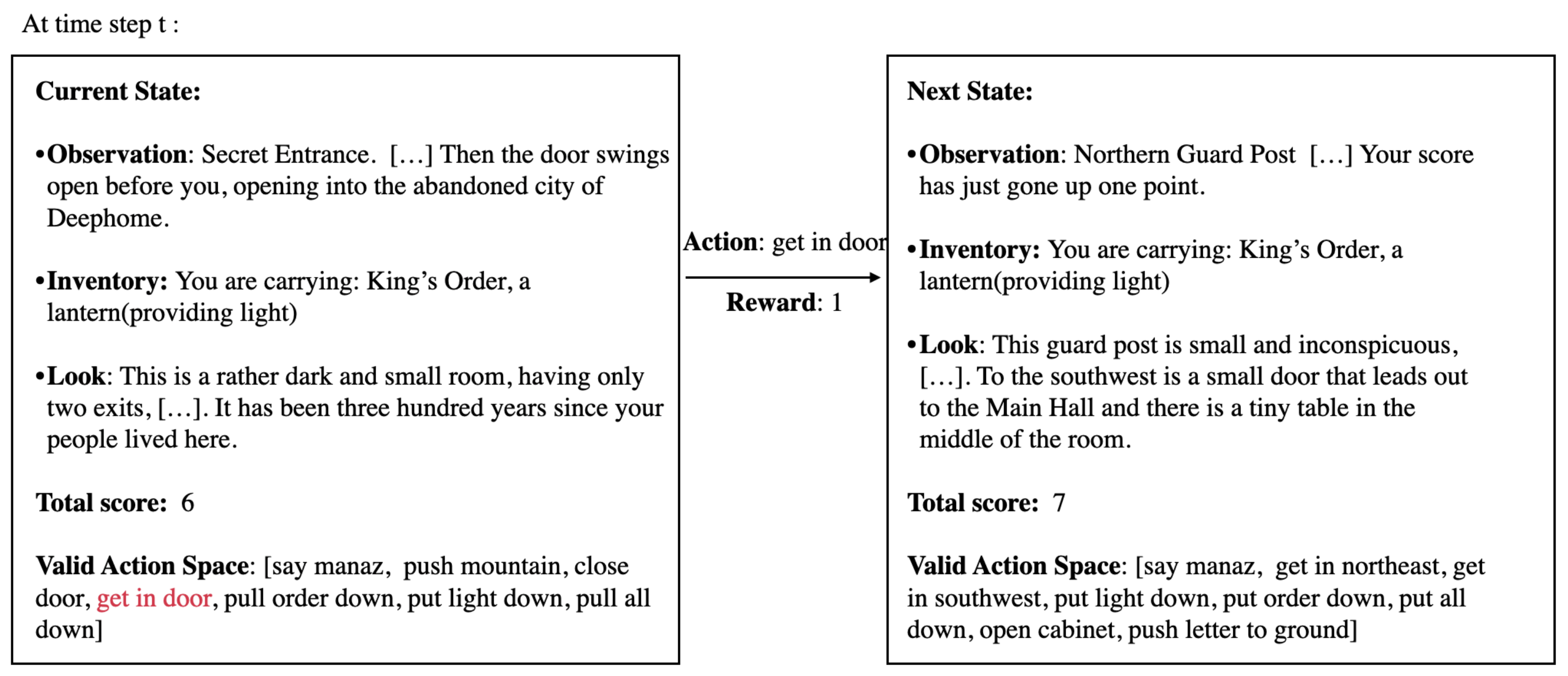
Text-based adventure games are a popular testbed for language-based reinforcement learning (RL). In previous work, deep Q-learning is often used as the learning agent. However, Q-learning algorithms are difficult to apply to complex real-world domains. We adapt the Soft-Actor-Critic (SAC) algorithm to the domain of text-based adventure games in this paper. To deal with sparse extrinsic rewards from the environment, we combine the SAC with a potential-based reward shaping technique to provide more informative (dense) reward signals to the RL agent. The SAC method achieves higher scores than the Q-learning methods on many games, with only half the number of training steps. Additionally, the reward shaping technique helps the agent learn the policy faster and improve the game score. Our findings show that the SAC algorithm is a well-suited approach for text-based games.

Introduction

Challenges of Text-based Adventure Game:

- The discrete action space is large and not fixed.
- Common-sense Reasoning and Knowledge Representation
- **Sparseness of rewards**, this problem is even more severe in text-based adventure games due to the large and context-dependent action space.

An example of the game Deephome:



At time step t :

Current State:

- **Observation**: Secret Entrance. […] Then the door swings open before you, opening into the abandoned city of Deephome.
- **Inventory:** You are carrying: King's Order, a lantern(providing light)
- **Look**: This is a rather dark and small room, having only two exits, […]. It has been three hundred years since your people lived here.

Total score: 6

Valid Action Space: [say manaz, push mountain, close door, get in door, pull order down, put light down, pull all down]

Action: get in door

Reward: 1

Next State:

- **Observation**: Northern Guard Post […] Your score has just gone up one point.
- **Inventory:** You are carrying: King's Order, a lantern(providing light)
- **Look**: This guard post is small and inconspicuous, […]. To the southwest is a small door that leads out to the Main Hall and there is a tiny table in the middle of the room.

Total score: 7

Valid Action Space: [say manaz, get in northeast, get in southwest, put light down, put order down, put all down, open cabinet, push letter to ground]

SAC for Discrete Action Spaces

Goal: We propose to use SAC as an alternative for text-based adventure games to overcome the drawbacks of deep Q-learning.

In the **critic part** [1], the targets for the Q-functions:

$$y(r,s',d) = r + \gamma(1-d)\left(\min_{i=1,2}\left(Q_{\hat{\theta}_i}(s')\right) - \alpha\log\left(\pi_\phi(s'_t)\right)\right), \tag{1}$$

The critic learns to minimize the distance between the target soft Q-function and the Q-approximation with stochastic gradients:

$$\nabla J_Q(\theta) = \nabla\mathbb{E}_{a\sim\pi(s),s\sim D}\frac{1}{B}\sum_{i=1,2}\left(Q_{\theta_i}(s) - y(r,s',d)\right)^2, \tag{2}$$

The **actor policy** update:

$$\nabla J_\pi(\phi) = \nabla\mathbb{E}_{s\sim D}\frac{1}{B}\left[\pi_t(s)^T[\alpha\log\pi_\phi(s) - \min_{i=1,2}(Q_{\theta_i}(s))]\right]. \tag{3}$$

The Potential-based Reward Shaping Method

Goal: We propose a variant of potential-based reward shaping to speed up the convergence.

The shaped function F at learning step:

$$F(s,a,s') = \gamma_r V(s') - V(s), \tag{4}$$

The new shaped reward $\hat{R}$:

$$\hat{R}(s,a) := R(s,a) + F(s,a,s'), \tag{5}$$

The soft value function:

$$V(s) = \pi(s)^T[Q_{\hat{\theta}_i}(s) - \alpha\log(\pi(s))], \tag{6}$$

Now, rewrite the target equation:

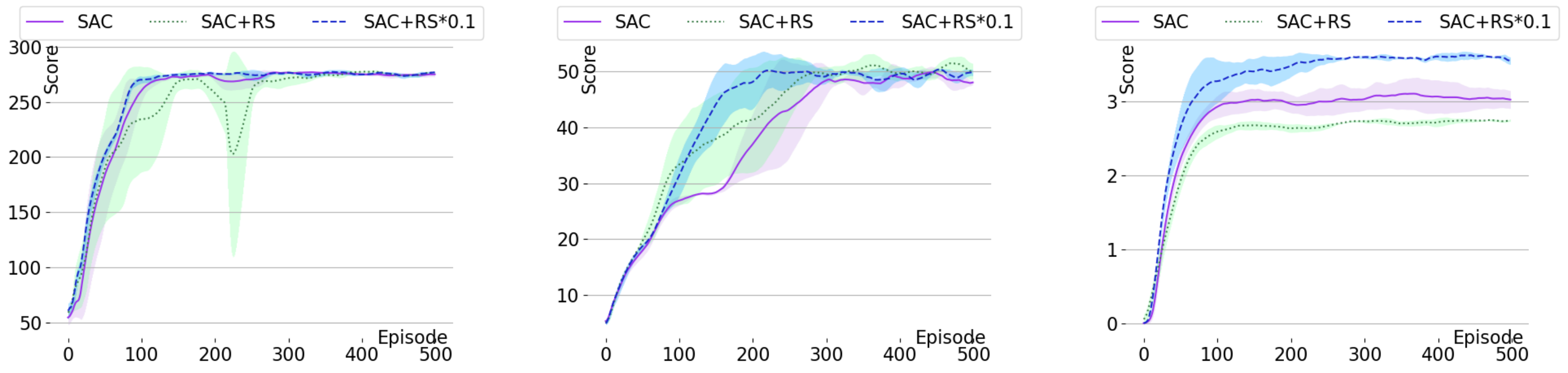
$$y(r,s',d) = [r + (\gamma_r V(s') - V(s))] + \gamma_r(1-d)V(s'). \tag{7}$$

Results with SAC

		Hausknecht *et al.* [2]			Yao *et al.* [3]	Ours
Game	Max	**RAND**	**DRRN**	**NAIL**	**DRRN**	**SAC**
adventureland	100	0	20.6	0	-	**24.8**
detective	360	113.7	197.8	136.9	**290**	274.8
pentari	70	0	27.2	0	26.5	**50.7**
balances	51	10	10	10	10	10
gold	100	0	0	3	-	**6.3**
jewel	90	0	1.6	1.6	-	**8.8**
deephome	300	1	1	13.3	**57**	48.1
karn	170	0	**2.1**	1.2	-	0.1
ludicorp	150	13.2	13.8	8.4	12.7	**15.1**
zork1	350	0	32.6	10.3	**39.4**	25.7
zork3	7	0.2	0.5	1.8	0.4	**3.0**
yomomma	35	0	0.4	0	-	**0.99**

- The average score of the **last** 100 episodes is shown for three repetitions of each game, and training on eight parallel environments.
- SAC aims to maximize the log probability, which can encourage the agent to explore uncertain states and converge faster.

Results SAC with Reward Shaping



(a) detective (b) deephome (c) zork3

- This figure compares the performance of the SAC agents with and without different reward-shaping variants, where shaded areas correspond to standard deviations.
- We find the reward shaping technique particularly advantageous for difficult games compared to possible games.

Limitations and Future Work

Limitations:

- The valid action spaces are often incomplete.
- The current RL agent needs a more robust semantic understanding.

Future Work:

- Generating accurate action spaces is crucial to improve agent performance.
- To incorporate semantic information in the agent and ensure it is used to predict the next action.

References

[1] Christodoulou, P.: Soft actor-critic for discrete action settings. arXiv preprint arXiv:1910.07207 (2019)

[2] Hausknecht, M., Ammanabrolu, P., Côté, M.A., Yuan, X.: Interactive fiction games: A colossal adventure. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7903–7910 (2020)

[3] Yao, S., Narasimhan, K., Hausknecht, M.: Reading and acting while blindfolded: The need for semantics in text game agents. pp. 3097–3102. Association for Computational Linguistics (2021)

weichen@cs.uni-kl.de/sophie.fellenz@rptu.de