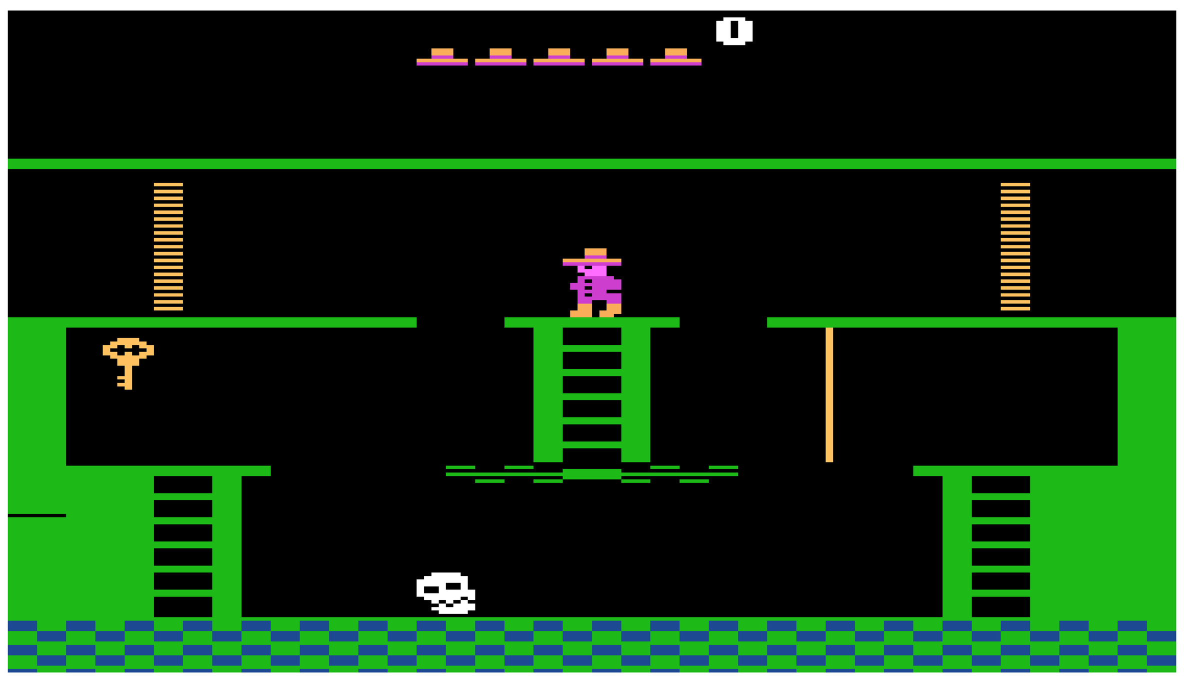
SDRL: Interpretable and Data-efficient Deep Reinforcement Learning Leveraging Symbolic Planning

Daoming Lyu[1], Fangkai Yang[2], Bo Liu[1], Steven Gustafson[3]
[1]Auburn University, Auburn, AL; [2]NVIDIA, Redmond, WA; [3]Maana Inc., Bellevue, WA

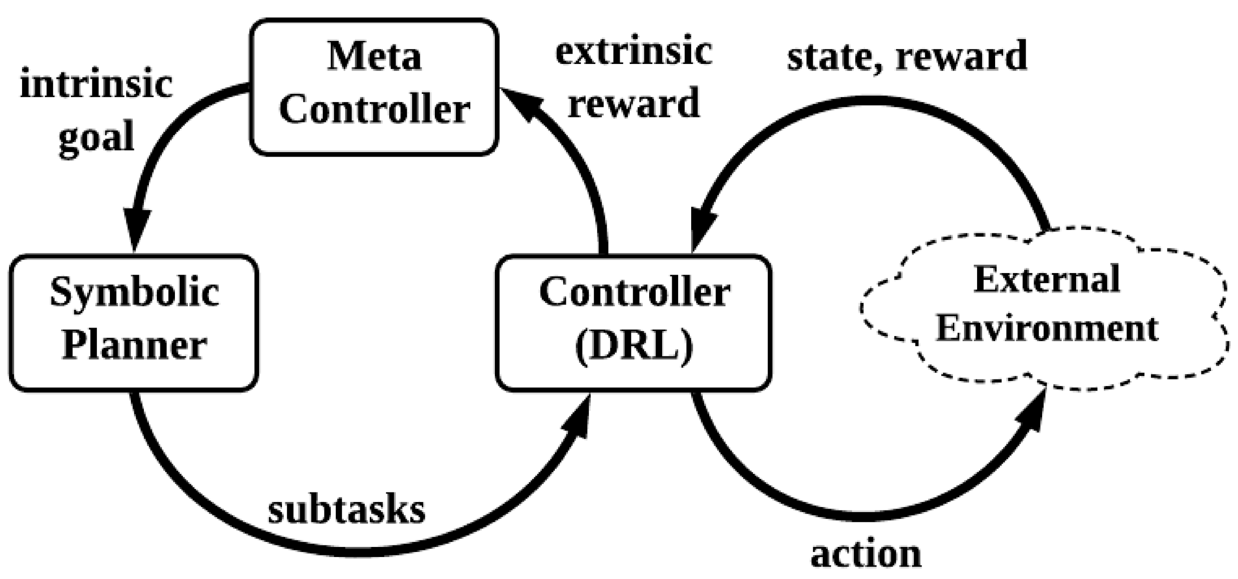
Problem

- Sequential decision-making with **long horizon action sequence** and **sparse reward** suffers from:
  - Poor data efficiency,
  - Lack of interpretability.
- Challenge: Montezuma's Revenge
  - The avatar: climbs down the ladder, jumps over a rotating skull, picks up a key (+100), goes back and uses the key to open the right door (+300).
  - Vanilla DQN achieves 0 score *(Mnih et al., 2015)*.



SDRL: Symbolic Deep Reinforcement Learning

- Goal:
  - Symbolic planning drives learning, improving **task-level interpretablility**.
  - DRL learns feasible subtasks, improving **data-efficiency**.
- Task decomposition.



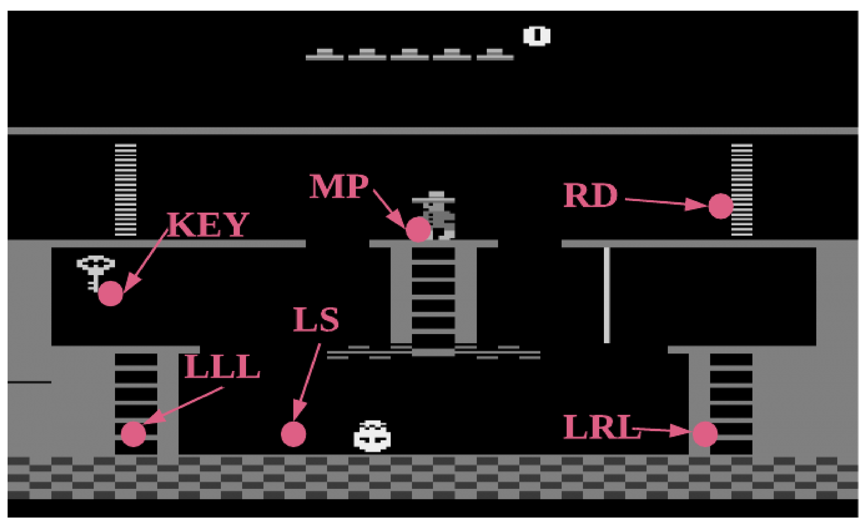
- **Symbolic Planner**: high-level symbolic planning based on intrinsic goal.
  - Intrinsic goal: a linear constraint on plan quality $quality \geq quality(\Pi_t)$, where Π_t is the plan at episode t.
  - Plan quality: a utility function that sums up the gain rewards of subtasks in a plan.
  - Mapping from symbolic transition to subtask.
- **Controller**: low-level policy control with DRL.
  - Intrinsic reward: pseudo-reward crafted by the human.
- **Meta-Controller**: subtask learning evaluation.
  - Extrinsic reward: a function about ϵ where ϵ is a criterion that measures the competence of the learned subpolicy for each subtask.
  - ϵ: success ratio (in our case).
  - Learnable subtask and unlearnable subtask.

Experimental Results

- Symbolic representation and predefined subtasks

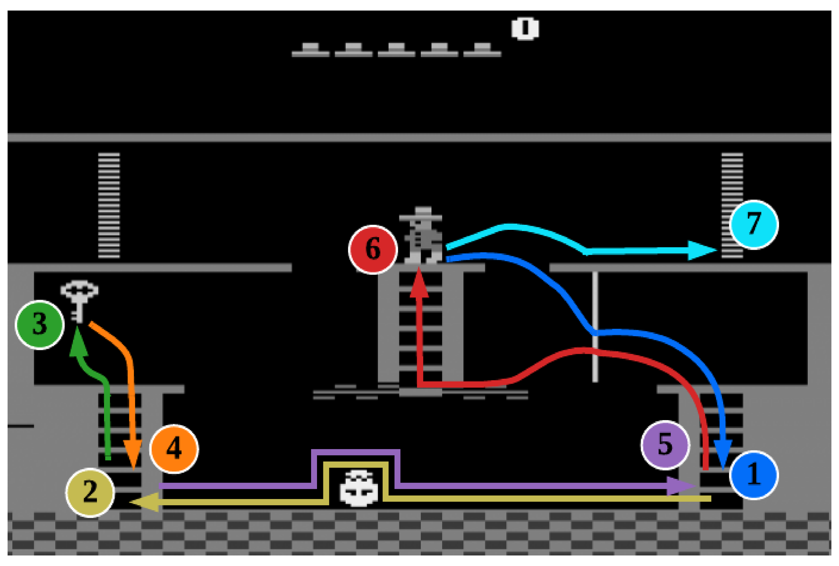
```
% object declaration
location(mp;rd;ls;lll;lrl;key).
% dynamic causal law declaration
move(L) causes loc=L if location(L).
move(L) causes cost=L+Z if rho((at(L1)),move(L))=Z,
           loc=L1,picked(key)=false.
move(L) causes cost=L+Z if rho((at(L1),picked(key)),
           move(L))=Z,loc=L1,picked(key)=true.
inertial loc. inertial quality.
% static causal law declaration
picked(key)=true if loc=key.
nonexecutable move(key) if picked(key).
default rho((at(L1)),move(L))=10.
default rho((at(L1),picked(key)),move(L))=10.
```

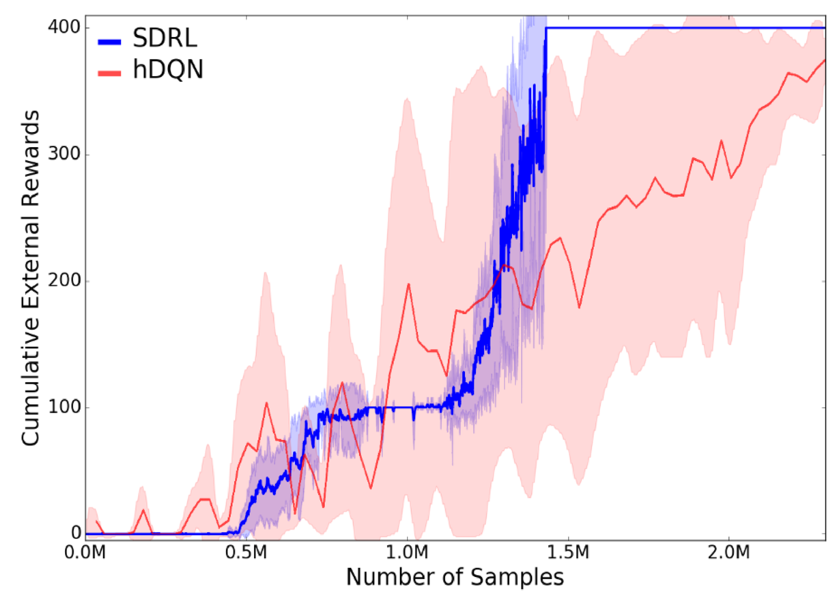
No.	subtask	policy learned	in optimal plan
1	MP to LRL, no key	✓	✓
2	LRL to LLL, no key	✓	✓
3	LLL to key, no key	✓	✓
4	key to LLL, with key	✓	✓
5	LLL to LRL, with or without key	✓	✓
6	LRL to MP, with or without key	✓	✓
7	MP to RD, with key	✓	✓
8	LRL to LS, with or without key	✓	
9	LS to key, with or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

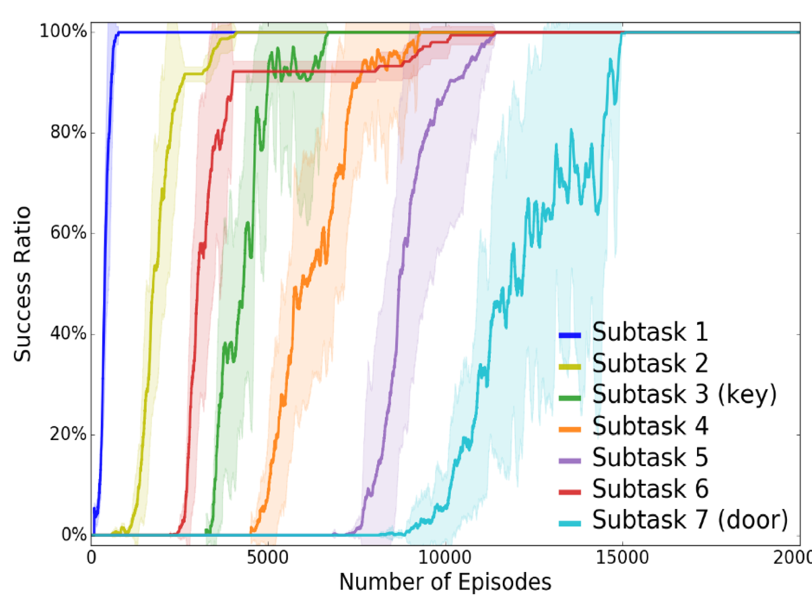


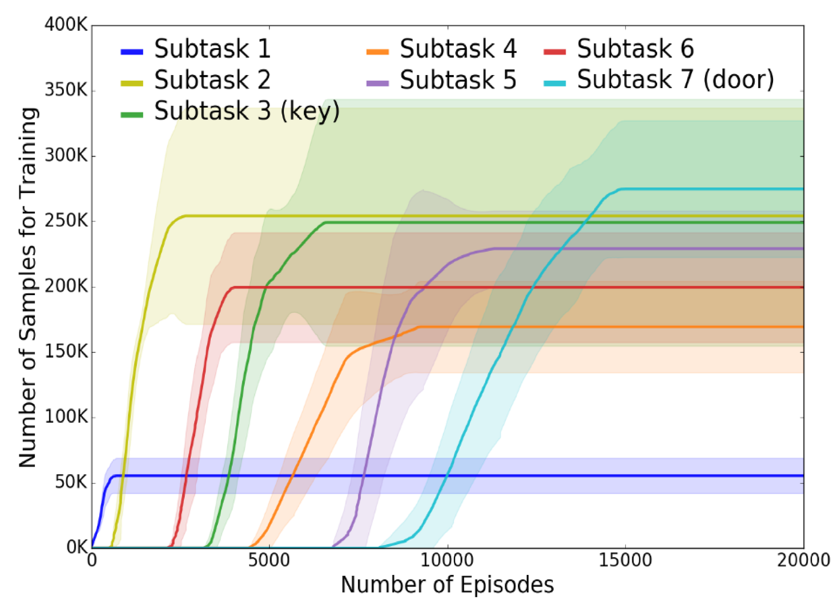
MP: middle platform
LRL: lower right ladder
LLL: lower left ladder
KEY: key
LS: left of rotating skull
RD: right door

- Final solution and learning curves









Conclusion

- We present the **SDRL** framework, and it is the first work on integrating symbolic planning with DRL that achieves both **task-level interpretability** and **data-efficiency** for decision-making.
- Future work will investigate on the transferability, and integration with automatic option discovery.

Reference

- 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*, pages 3675–3683. (**our baseline**)
- 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.
- Yang, F., Lyu, D., Liu, B., and Gustafson, S. (2018). Peorl: Integrating symbolic planning and hierarchical reinforcement learning for robust decision-making. In *International Joint Conference of Artificial Intelligence* (IJCAI).