Rank the Episodes: A Simple Approach for Exploration in Procedurally-Generated Environments

Daochen Zha†, Wenye Ma‡, Lei Yuan‡, Xia Hu†, Ji Liu‡

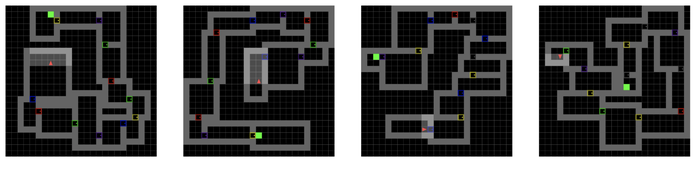
†*Department of Computer Science and Engineering, Texas A&M University*
‡*AI Platform, Kwai Inc.*

Background: Exploration is an Open Challenge for Reinforcement Learning

- **What is exploration?** Exploration is the ability of the agents to discover novel states in the environments.
- **Why we need exploration?** Learning a good policy with reinforcement learning relies on the ability of discovering the novel states in the first place. However, in many environments, the rewards are sparse and cannot be easily discovered.
- **How we can encourage exploration?** The most popular method for encouraging exploration is to give intrinsic rewards based on uncertainty, e.g., count-based exploration, curiosity driven exploration, etc.

Exploration Needs to Generalize

- **Over-fitting Issue.** The agent is usually trained and tested in the save environment, leading to over-fitting.
- **Procedurally-Generated Environments.** To address this issue, some procedurally-generated environments are proposed, where a different environment is generated in each episode. For example, the following figure presents the mazes of MiniGrid environments in four diffrent episodes.



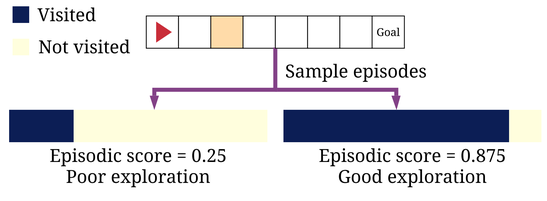
- **Why challenging?** The agent needs to learn from diverse environments and generalize to unseen environments.

Can Existing Methods Generalize?

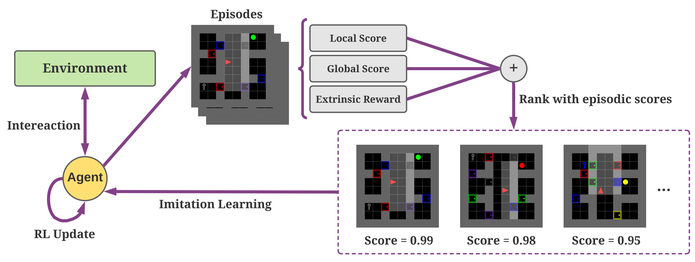
- **Motivating Example:** We take count-based exploration as an example and consider a 1D Maze, where the agent (red) needs to move right in each step to reach a goal. We visualize the intrinsic rewards for singleton (i.e., same environment in each episode) and procedural-generated setting (the observation of the third block is randomly sampled from 10 values).

- **Count-Based Exploration is Less Effective.** The agent may get stuck in the third block. This is because visiting a novel state does not necessarily mean a good exploration behavior in procedurally-generated setting.

Episode-Level Exploration



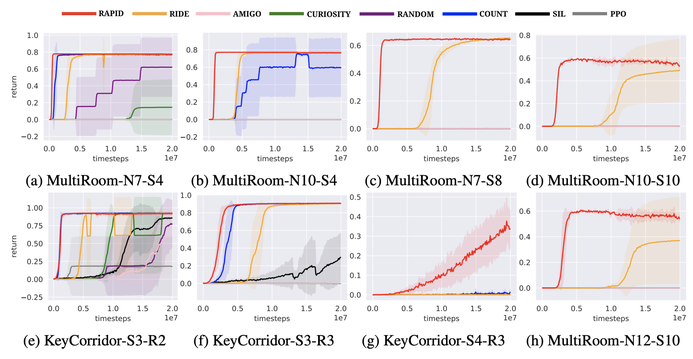
- **Episodic Score:** Using the the coverage rate (the number of distinct states divided by the total number of states), we can easily distinguish good exploration in episode-level.



We introduce **Ra**nk the E**pi**so**d**es (RAPID) algorithm.

- **Step 1 Rank**: Collect episodes and give episodic exploration scores. Store the good episodes in a buffer
- **Step 2 Behavior Cloning**: Perform imitation learning to the episodes in the buffer.

New SOTA on MiniGrid



(a) MultiRoom-N7-S4 (b) MultiRoom-N10-S4 (c) MultiRoom-N7-S8 (d) MultiRoom-N10-S10

(e) KeyCorridor-S3-R2 (f) KeyCorridor-S3-R3 (g) KeyCorridor-S4-R3 (h) MultiRoom-N12-S10

Paper **Code**