Episodic Novelty Through Temporal Distance

Yuhua Jiang*, Qihan Liu***,** Yiqin Yang, Xiaoteng Ma, DianyuZhong
Jun Yang, Bin Liang, Bo Xu, Chongjie Zhang, Qianchuan Zhao

2024.12.15

Introduction

- Reinforcement Learning (RL) has achieved great success.
  - Given a MDP $\langle \mathcal{S}, \mathcal{A}, P, r, \mu_S, \gamma \rangle$
  - Optimize a policy π to maximize the cumulative reward:

$$\max_{\pi} \mathbb{E}_{s_0 \sim \mu_S, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t)} \left[\sum \gamma^t r(s_t, a_t, s_{t+1}) \right]$$

  - Sparse reward is the challenge.
    - Hard exploration.

Introduction

- Exploration in Sparse Reward Setting
  - Intrinsic Motivation

Sparse Reward

Intrinsic Reward (Bonus)

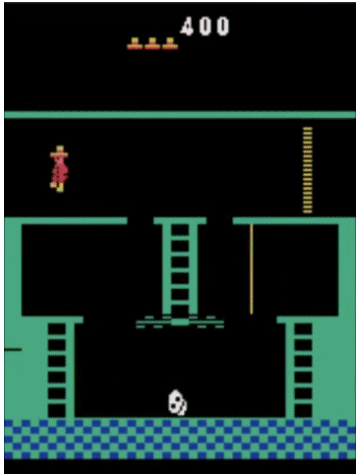
$$r(s_t, a_t, s_{t+1}) = \boxed{r_t^e} + \beta \cdot \boxed{b_t}$$

- Encouraging the agent to actively **explore**,

  especially states that **are novel or rare**.

Introduction

- Intrinsic rewards in singleton MDPs has been well-studied
  - Singleton MDP
    - Agent is initialized in the same environment at each episode



Montezuma's Revenge

- Global (lifelong) intrinsic reward performs well
  - Rely on lifelong experiences.
- ICM, RND, …

Method	Intrinsic Bonus: $b_{\text{Method}}(s_t)$
ICM	$\|\|\hat{\phi}(s_t) - \phi(s_t)\|\|_2^2$
RND	$\|\|f(s_t) - \bar{f}(s_t)\|\|_2^2$

Introduction

- Contextual MDP (CMDP)
  - Different episodes correspond to different environments but share structure.
    - Global intrinsic reward fail.
    - Episodic intrinsic reward is preferable.
      - Rely on experiences from current episode.

Results From [1]

$$b_{\text{global}}(s) = \frac{1}{\sqrt{N(\psi(s))}}, \quad b_{\text{episodic}}(s) = \mathbb{I}[N_e(\psi(s)) = 1]$$

Environment	$\lvert\mathcal{C}\rvert$	Global	Episodic
MultiRoom	1	0.99 ± 0.00	0.83 ± 0.23
MultiRoom	3	0.59 ± 0.32	0.92 ± 0.13
MultiRoom	5	0.23 ± 0.39	0.98 ± 0.02
MultiRoom	10	0.02 ± 0.06	0.78 ± 0.17
MultiRoom	∞	0.00 ± 0.00	0.87 ± 0.10

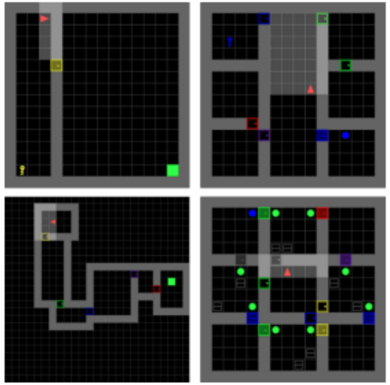
[1] *Henaff, Mikael, et,al*. A Study of Global and Episodic Bonuses for Exploration in Contextual MDPs. ICML 23.

Introduction

- Definitions of CMDP

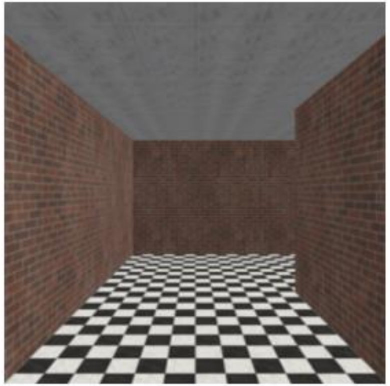
  $$(\mathcal{S}, \mathcal{A}, \mathcal{C}, P, r, \mu_C, \mu_S, \gamma)$$

  - Introduce context (C)
    - Dynamics $P : \mathcal{S} \times \mathcal{A} \times \mathcal{C} \to \Delta(\mathcal{S})$
    - Initial states $\mu_S(\cdot|c)$
- Examples of CMDP
  - Procedurally generated environments



(a) MiniGrid

(b) Crafter



(c) MiniWorld

Related Work

Method	Intrinsic Bonus: $b_{\text{Method}}(s_t)$	Episodic Bonus Category
AGAC	$D_{\mathrm{KL}}\left(\pi\left(\cdot \mid s_t\right) \\| \pi_{\mathrm{adv}}\left(\cdot \mid s_t\right)\right)+\beta \cdot \frac{1}{\sqrt{N_e(s_{t+1})}}$	Count
RIDE	$\\|\phi\left(s_{t+1}\right)-\phi\left(s_t\right)\\|_2 \cdot \frac{1}{\sqrt{N_e(s_t)}}$	Count
NovelD	$[b_{\mathrm{RND}}(s_{t+1})-b_{\mathrm{RND}}(s_t)]_{+} \cdot \mathbb{I}[N_e(s_t)=1]$	Count
NGU	$b_{\mathrm{RND}}(s_t) \cdot \frac{1}{\left(\sqrt{\sum_{\phi_i \in N_k} K(\phi(s_t), \phi_i)+c}\right)}$	Similarity
E3B	$\phi\left(s_t\right)^{\top}\left[\sum_{i=0}^{t-1} \phi\left(s_i\right) \phi\left(s_i\right)^{\top}+\lambda I\right]^{-1} \phi\left(s_t\right)$	Similarity
EC	$\alpha(\beta-F\{C(s_i, s_t)\}_{i \in \|M\|})$	Similarity
DEIR	$\min_{i \in \|M\|}\left\{\frac{\|\|\phi(s_i), \phi(s_t)\|\|^2}{\|\|\phi_{\mathrm{rnn}}(s_i), \phi_{\mathrm{rnn}}(s_t)\|\|}\right\}$	Similarity

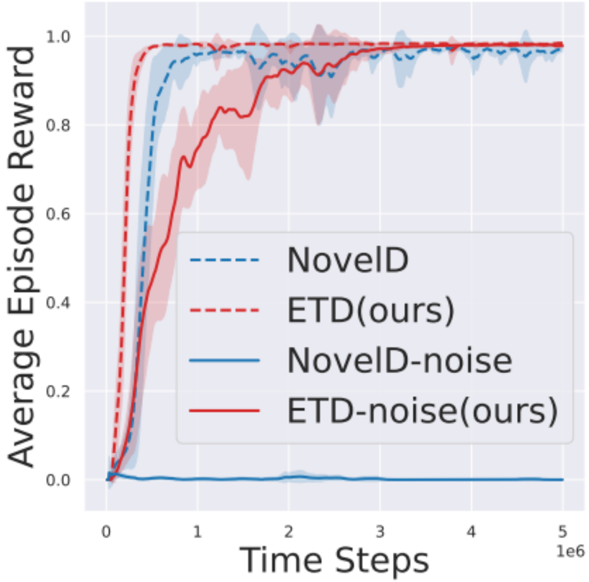
A summary of recent research on CMDP exploration

Limitations of Current Episodic Bonus

- Count-based
  - Hard to scale large state space
  - Example
    - NovelD:

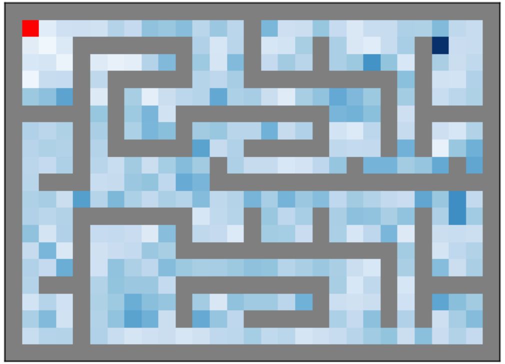
$$[b_{\text{RND}}(s_{t+1}) - b_{\text{RND}}(s_t)]_+ \cdot \mathbb{I}[N_e(s_t) = 1]$$

Add Gaussian noise to the states.

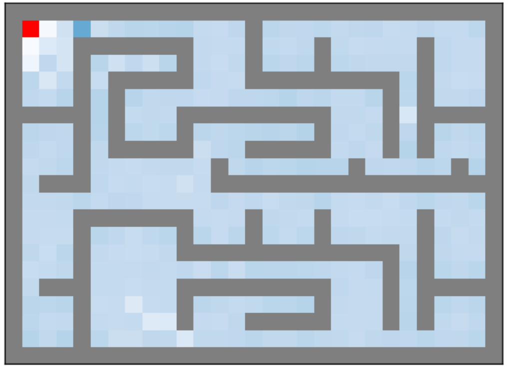


Limitations of Current Episodic Bonus

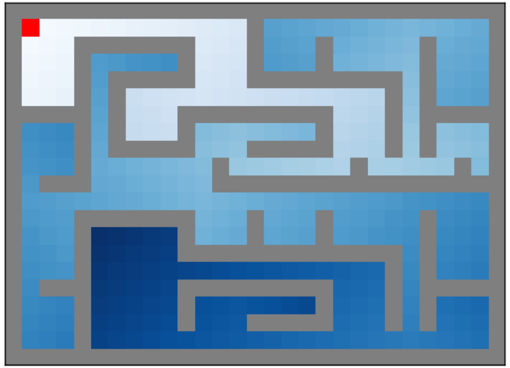
- Similar-based
  - Current similarity measurement is not suitable.



Euclidean distance under inverse dynamics representation (e.g., NGU, E3B)



Reachability probability (e.g., EC, DEIR)



Temporal Distance (Expected)
The focus of this work

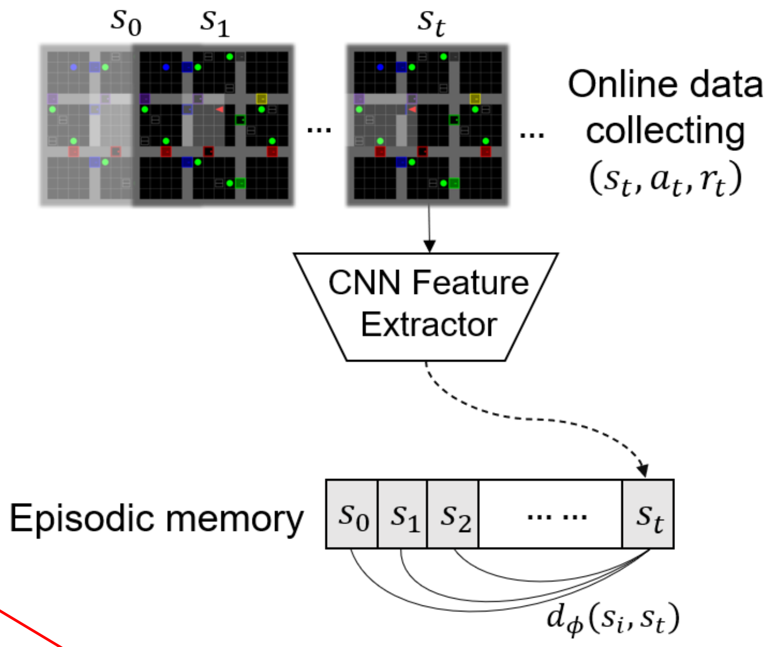
Our Work

- Consider CMDP sparse reward problem
  - Using temporal distance as a metric for state similarity
  - Design a new episodic bonus
    - Encourage agents to explore states that are temporally distant from their episodic history.

  - We named our method, ETD (**E**pisodic Novelty Through **T**emporal **D**istance)

Our Proposed

- Assume we have already obtained the temporal distance function:
  - $d_\phi(x, y)$: $\mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$
- Episodic bonus defined as:
  - The minimum temporal distance between the current state and all states in the episode history.



Prevent repeated state visits.

Avoid staying in areas with easy-to-reach.

$$b_t = \min_{i \in [0,t)} d_\phi(s_i, s_t)$$

Temporal Distance Definitions

- Temporal distance
  - Probability of reaching from x to y

$$p_\gamma^\pi(s^f = y | s_0 = x) = (1-\gamma) \sum_{k=0}^{\infty} \gamma^k p^\pi(s_k = y | s_0 = x).$$

  - Propose to use Successor Distance [2]

$$\boxed{d_{\mathrm{SD}}^\pi(x, y) = \log\left(\frac{p_\gamma^\pi(s_f = y | s_0 = y)}{p_\gamma^\pi(s_f = y | s_0 = x)}\right)}$$

$-\log$ (from x to y – from y to y)

$\approx$ Hitting time from x to y

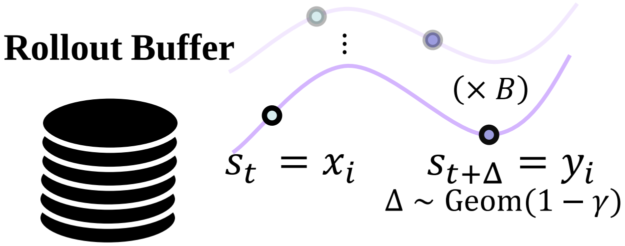
Satisfy quasimetric

(Positivity, Identity, Triangle Inequality)

[2] *Vivek Myers, et, al*. **Learning Temporal Distances: Contrastive Successor Features Can Provide a Metric Structure for Decision Making**. ICML, 2024.

Temporal Distance Learning

1. Sampling positive pairs $\{x_i, y_i\}_{i=1}^{B}$



2. Specially parameterized energy function

$$\underbrace{f_{(\psi,\phi)}(x,y)}_{\substack{\texttt{Energy function} \\ \mathcal{S}\times\mathcal{S}\to\mathbb{R}}} := \underbrace{c_\psi(y)}_{\substack{\texttt{Potential function} \\ \mathcal{S}\to\mathbb{R}}} - \underbrace{\boxed{d_\phi(x,y)}}_{\substack{\texttt{Quasimetric function} \\ \mathcal{S}\times\mathcal{S}\to\mathbb{R}}}$$

3. Optimize the symmetric InfoNCE loss

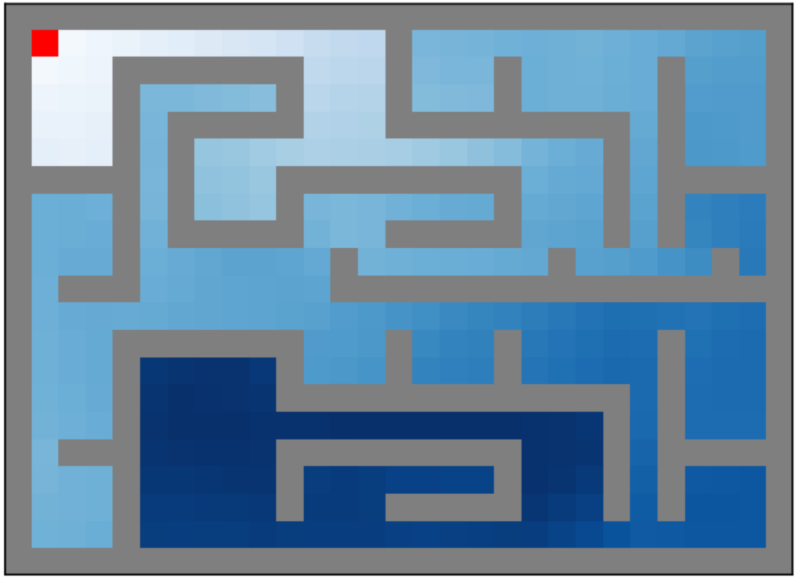
$$\mathcal{L}_{(\phi,\psi)} = \sum_{i=1}^{B}\left[\log\frac{\exp(f_{(\phi,\psi)}(x_i,y_i))}{\sum_{j=1}^{B}\exp(f_{(\phi,\psi)}(x_i,y_j))} + \log\frac{\exp(f_{(\phi,\psi)}(x_i,y_i))}{\sum_{j=1}^{B}\exp(f_{(\phi,\psi)}(x_j,y_i))}\right]$$

Classify (teal) from (beige) for each row and column.

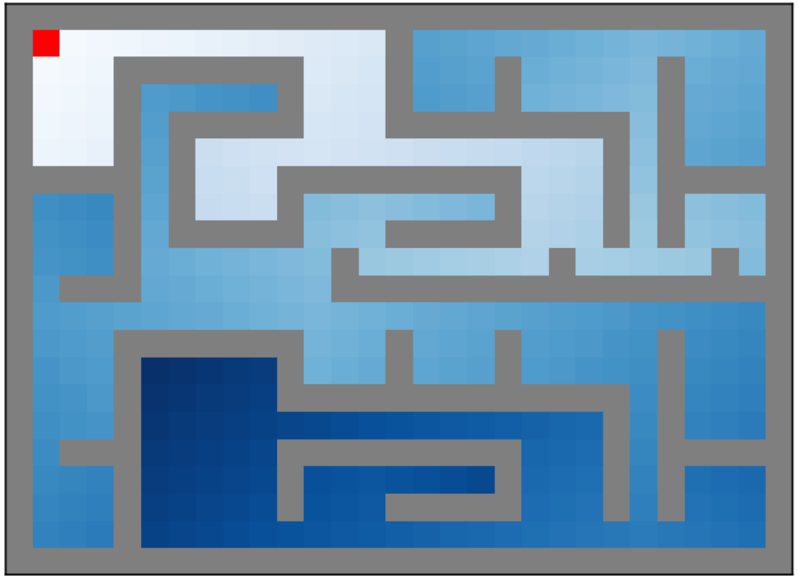
	y_1	y_2	$\cdots$	y_B
x_1	$f(x_1,y_1)$	$f(x_1,y_2)$	$\cdots$	$f(x_1,y_B)$
x_2	$f(x_2,y_1)$	$f(x_2,y_2)$	$\cdots$	$f(x_2,y_B)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_B	$f(x_B,y_1)$	$f(x_B,y_2)$	$\cdots$	$f(x_B,y_B)$

$d_\phi^*(x,y) \approx$ temporal distance from x to y

Temporal Distance Experiments

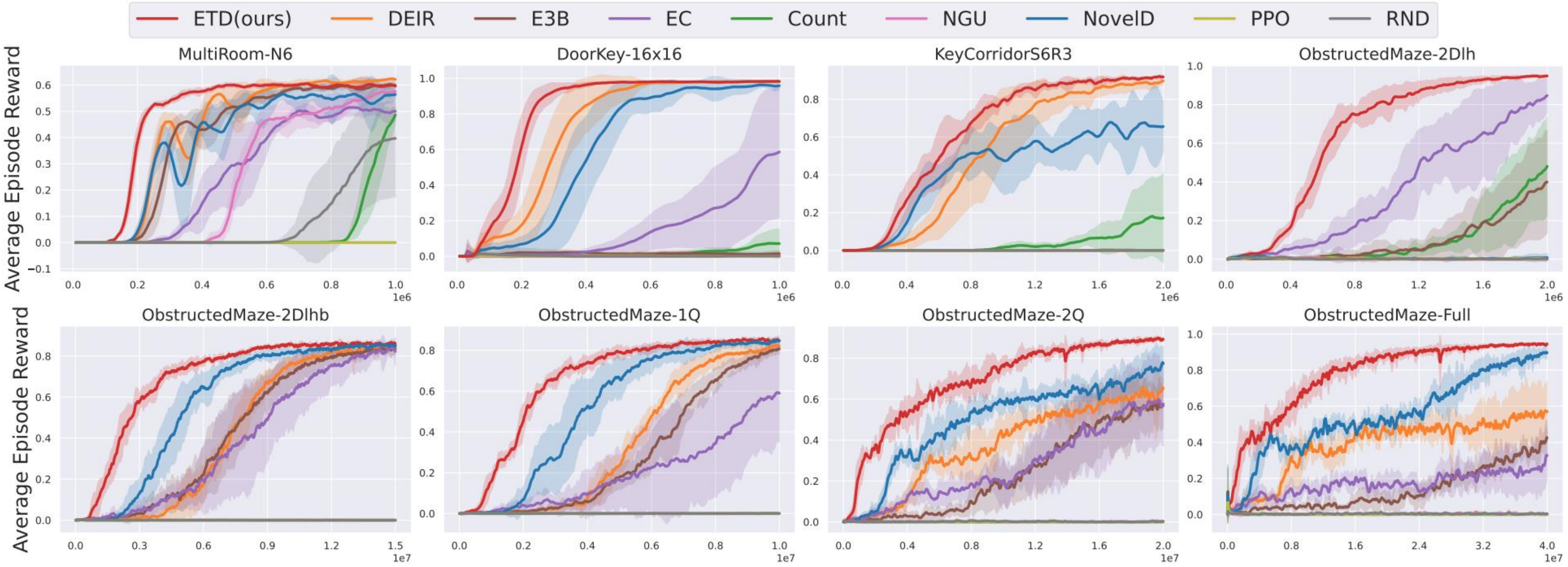


Ours

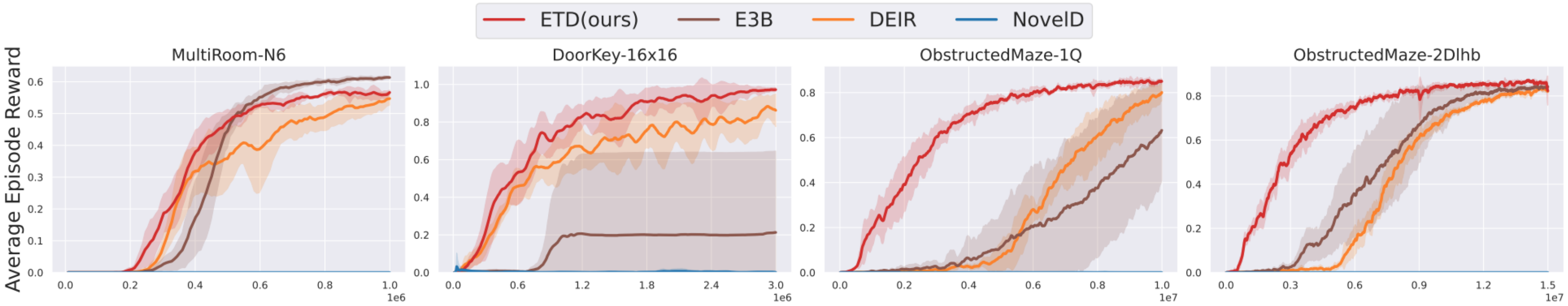


Ground Truth

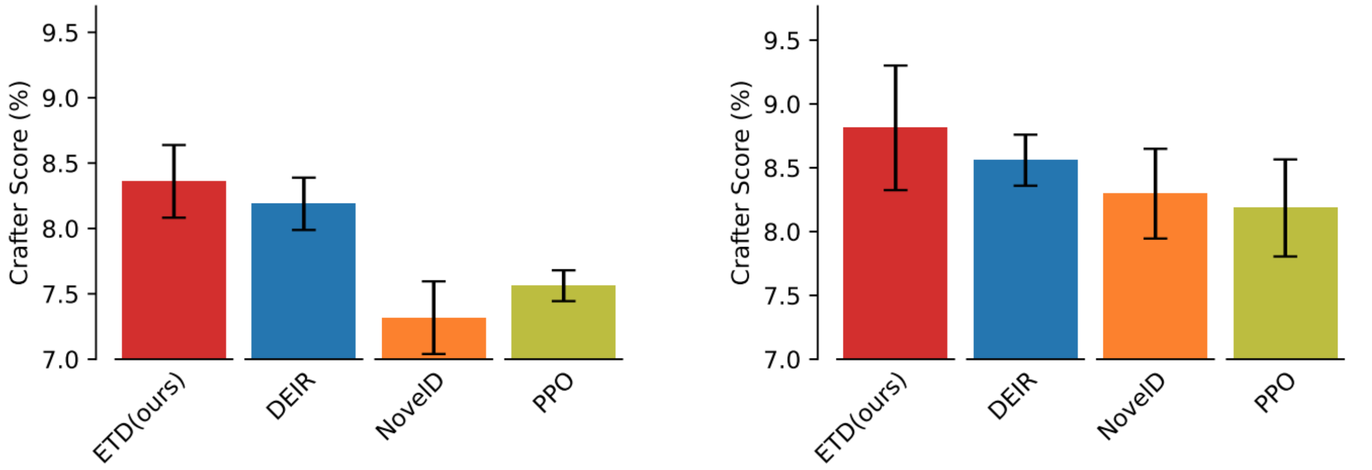
Performance on Exploration Tasks： MiniGrid



Experiments in MiniGrid with noise

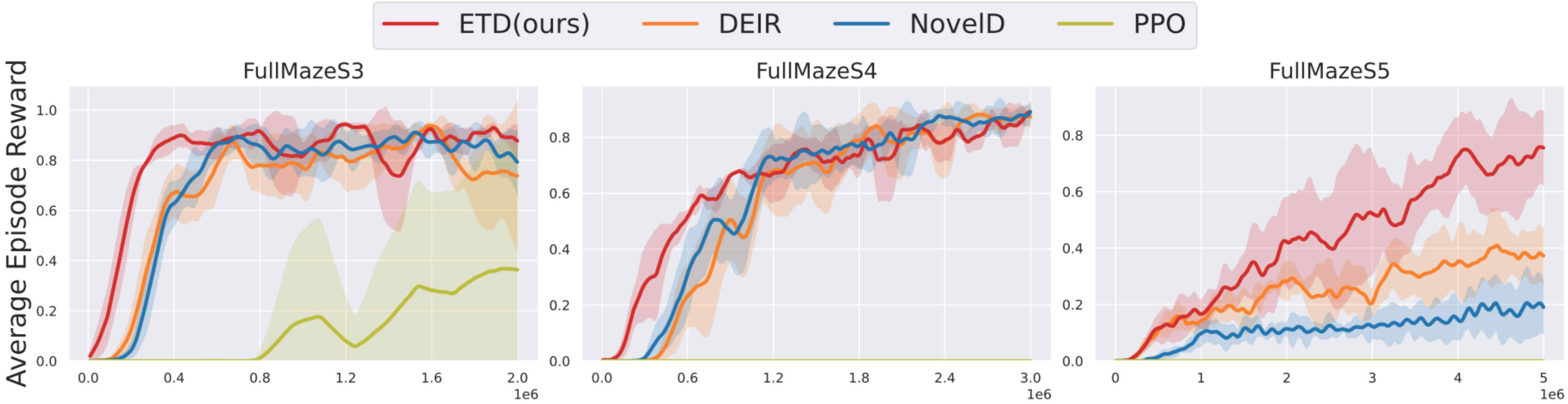


Experiments in Pixel-based Tasks: Crafter



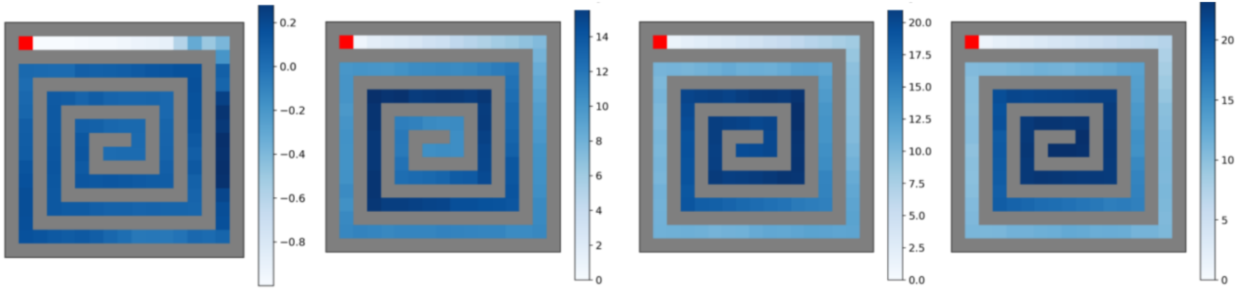
(a) Crafter (b) Crafter w.o life reward

Experiments in Pixel-based Tasks: MiniWorld



Ablations of energy functions

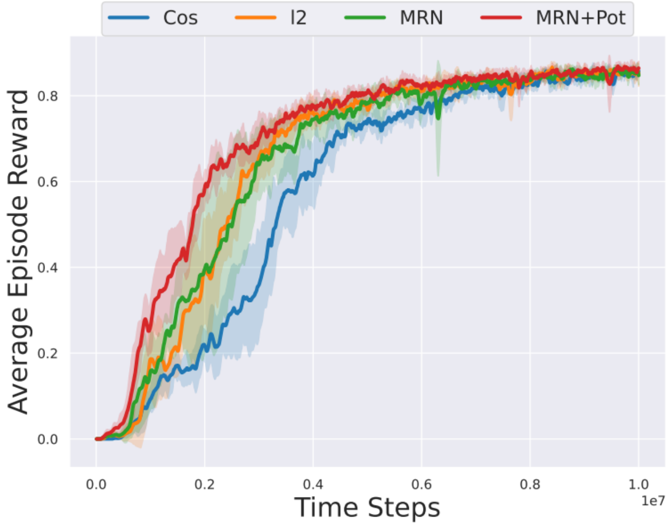
Cos l2 MRN MRN+POT



$$f_{\phi,\cos}(x,y) = \frac{\langle \phi(x), \phi(y) \rangle}{\|\phi(x)\|_2 \|\phi(y)\|_2},$$
$$f_{\phi,\mathrm{l}_2}(x,y) = -\|\phi(x) - \phi(y)\|_2,$$
$$f_{\phi,\mathrm{MRN}}(x,y) = -d_\phi(x,y),$$
$$f_{\phi,\psi,\mathrm{MRN+POT}}(x,y) = \psi(y) - d_\phi(x,y).$$

Both indicates: MRN+POT > MRN = l2 > Cos

The more closely with the temporal distance,
The higher the exploration efficiency.



Takeaways

- Episodic Bonus is crucial for CMDPs.
- Using temporal distance as a similarity metric for bonus design can significantly improve exploration efficiency.
- Code is available at https://github.com/Jackory/ETD.

Thanks!

Code is available at https://github.com/Jackory/ETD.