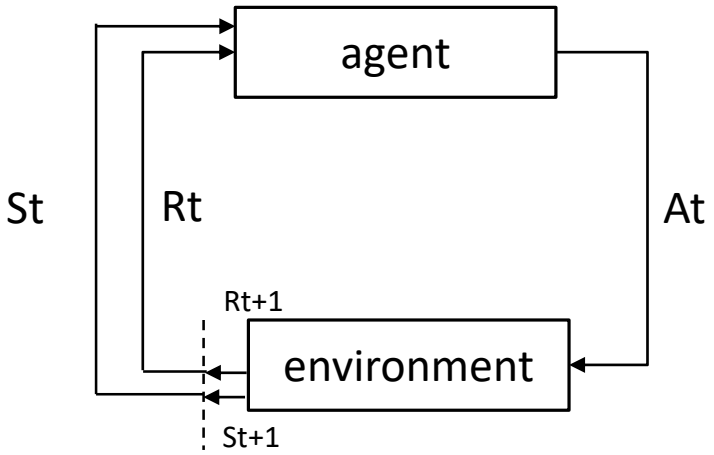
南京航空航天大學
Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Causality-driven Hierarchical Structure Discovery for Reinforcement Learning

NeurIPS | 2022

Reinforcement Learning



⟨S, A, R, S'⟩

1. decision-making
2. trial and error
3. interaction

4. value-based/policy based
5. model-based/model-free

hierarchical reinforcement learning

Hierarchical Reinforcement Learning (HRL) decomposes a long-horizon reinforcement learning task into a hierarchy of subproblems or subtasks such that a higher-level policy learns to perform the task by choosing optimal subtasks as the higher-level actions.

Example:

Task: Autonomous Driving
Subproblem 1: Trajectory Planning
Subtask 1: Lane Keeping
Subtask 2: Lane Changing
Subproblem 2: Obstacle Avoidance
Subtask 1: Obstacle Detection
Subtask 2: Collision Avoidance

Subgoal-based MDP:
Introduce the subgoal space G to extend the framework of MDP
(S, G, A, T, r, γ)

- G: subgoal space
- r(s, g, a, s′) : goal-reaching reward function that indicates whether the agent achieves subgoal g in transition (s, a, s′).

The proposed framework aims to pre-train the subgoal-based policy for downstream sparse reward tasks.

hierarchical reinforcement learning

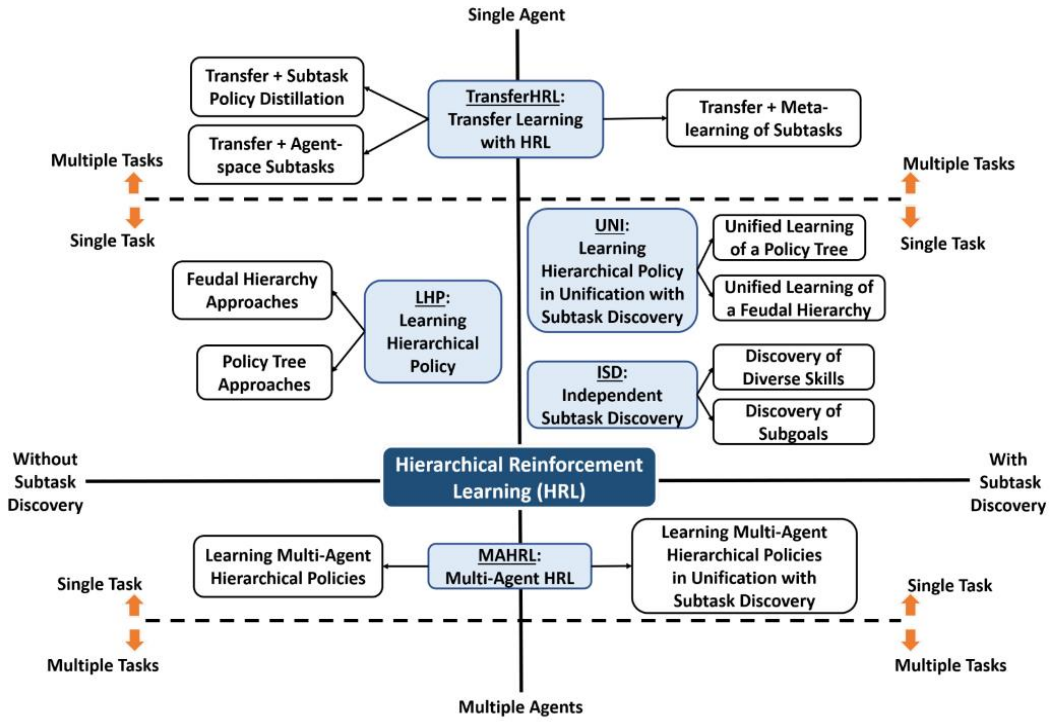


Fig. 3. **A taxonomy of HRL approaches**. The approaches are arranged along the following three dimensions: with or without subtask discovery, for single agent or multiple agents, and for single task or multiple tasks.

Advantages: efficient for decision making; avoiding complicated overall task.

Limitations: hard to match the true environment; time-consuming

Can we design a hierarchical system that can understand inner relationships between the variables?

Causality and Causality Discovery

Causality refers to the relationship between cause and effect, a idea that an event (the cause) leads to a subsequent event (the effect).

Characteristic of causality:

- the cause must happen before the effect
- The cause must be sufficient for the effect to occur
- can be determined through experiments or observational studies

Causal discovery is the process of identifying causal relationships between variables in a system

- Bayesian networks
- structural equation models
- causal inference algorithms

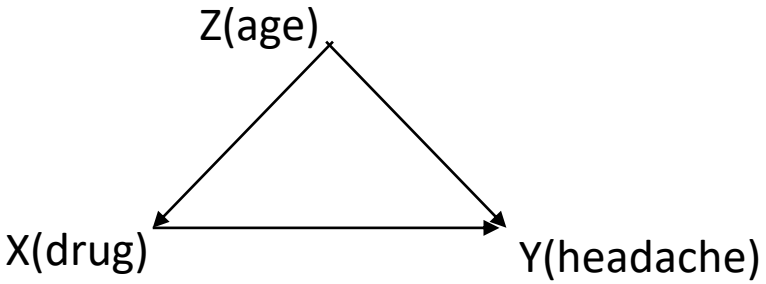
Structural Causal Model (SCM)

Structural Causal Model (SCM)

represent the causal relationships between variables in a system, where the nodes represent the variables in the system and the edges represent the causal relationships between them。

●Processes

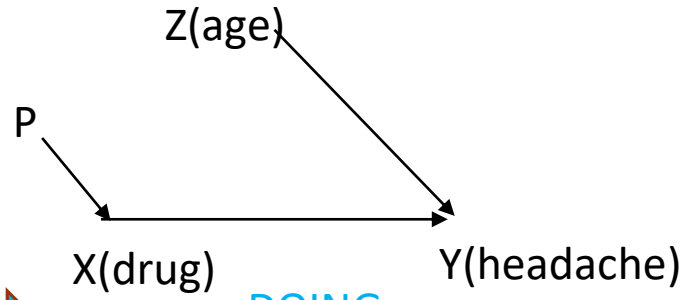
Drug ← fD (Age, UD)
Headache ← fH(Drug, Age, UH)



SEEING

●Intervention

Drug ← P
Headache ← fH (Drug, Age, UH)



DOING

(X, U, F, P(u)),

- X : endogenous variables
- U : exogenous variables,
- F : functions determining X, for each Xi,
  Xi :=$f_i(X_{pa(i,c)}, U_i)$
- P(u) is a distribution over U

Methodology

- How to discover causality using SCM
- How to get intervention data
- How to bridge causality discovery and subgoal hierarchy construction?

Methodology

How to discover causality using SCM

Causality is discovered within adjacent steps.

$$f_i(X_{pa(i,c)}, N_i) \longrightarrow f_i(X_{pa(i,c),t}, N_i)$$

The paper aims to learn the causality from transition data of adjacent steps in the agent’s trajectory

只研究相邻步数间的因果关系

How to get intervention learning data

Intervention data is obtained through subgoal-based policy

- 对于可控变量X_i，通过均匀分布采样得到x_i，并作为subgoal加入到策略中
- 定位后续轨迹直到该变量值变化为采样值x_i，在相邻步骤中获取该变量的干预数据

Methodology

Xi :=$f_i(X_{pa(i,c)}, U_i)$

Define:

structural parameters η： model causal graph C(MxM tensor)

Functional parameters θ: model function f

$\sigma(\eta_{ij})$ represents the probability that Xj is a direct cause of Xi(σ is sigmoid)

θi are the parameters of Xi's conditional probability function fi given Xi's parent variable set Xpa(i,C).

F :function parameters θ’ s training times in one iteration,

Q: is structural parameters η’ s training times in one iteration,

K: is the sampling times to estimate the gradient of η.

Methodology

Algorithm 2 CausalityDiscovery

Parameter Structural parameters $\eta \in R^{M\times M}$; Functional parameters θ; Intervention data set D_I; Intervention variable set S_{IV};

while T times iteration **do**
  while $f_{idx} < Fs$ **do**
    for variable $i = 0$ to $Size(S_{IV})$ **do**
      $X \sim D_I[i]$
      $C \sim Ber((\sigma(\eta)))$
      $L = -\log P(X|C;\theta_i)$
      $\theta_i \leftarrow Adam(\theta_i, \nabla_{\theta_i} L)$
    end for
    $f_{idx} = f_{idx} + 1$
  end while
  while $q_{idx} < Qs$ **do**
    for variable $j = 0$ to $Size(S_{IV})$ **do**
      while $k < K$ **do**
        $X \sim D_I[j]$
        $C^{(k)} \sim Ber((\sigma(\eta)))$
        $L_C^{(k)} = -\log P(X|C;\theta)$
        $k = k + 1$
      end while
      $g_{:,j} = \sum_k (\sigma(\eta_{:,j}) - c_{:j}^{(k)}) \left(\frac{L_C^{(k)}}{\sum_k L_C^{(k)}}\right)$
      $\eta_j \leftarrow \eta_j + g_j$
    end for
    $q_{idx} = q_{idx} + 1$
  end while
end while
return $C = \{\mathbf{1}_{\sigma(\eta_{ij})>0.8}\}$

Phase 1(function learning):

- Bernouli sample from C(intervention data)
- Maximize the likelihood of step one data

Phase 2(graph learning):

- estimate the gradient of η

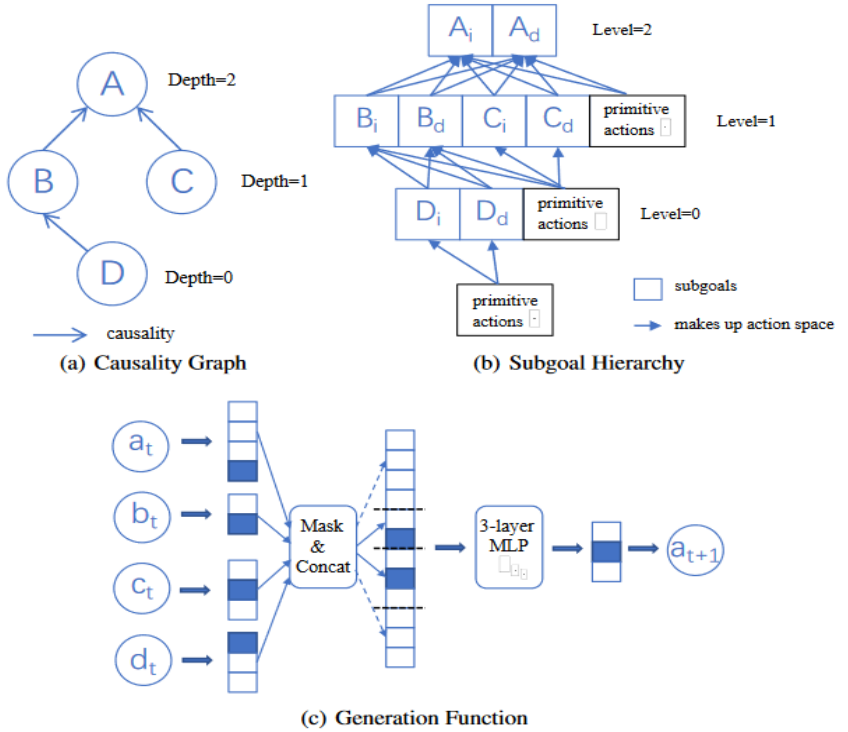
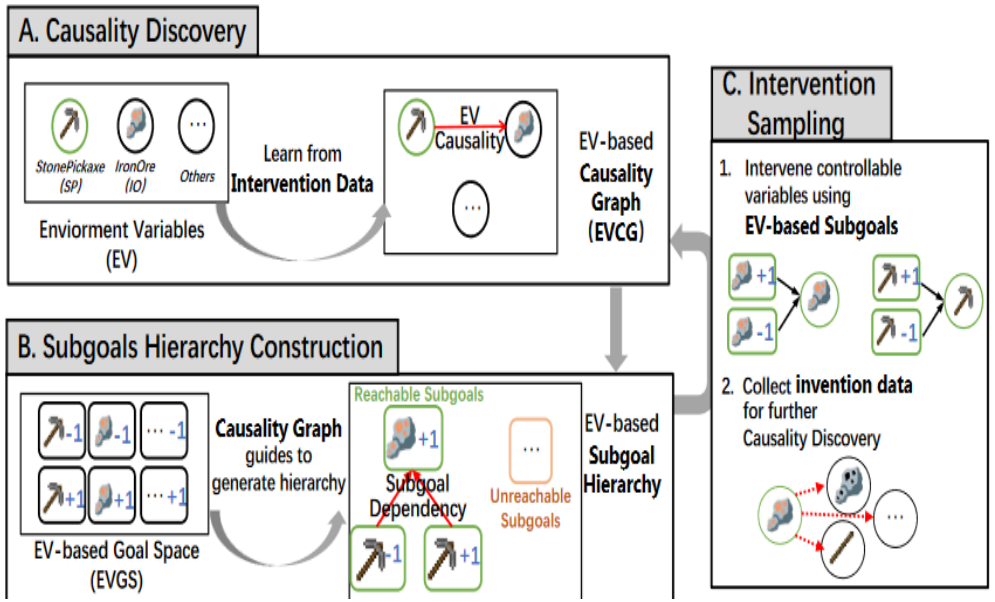


Figure 6: An implementation example. (a): A causality graph with four variables A, B, C, D. (b): Implementation of the subgoal hierarchy based on the causality graph in (a). The arrows pointing to the subgoal indicate the subgoal's action space. For example, subgoal B_i's action space consists of B's parent variable D's subgoals and primitive action space. (c): Implementation of variable A's generation function f_{θ_A}.

Methodology

How to bridge causality discovery and subgoal hierarchy construction?

Two Characteristic;
Three Periods;



Goal space design:

Two changes: increase and decrease

$$F_i(x_{i,t}, x_{i,t+1}) = \mathbf{1}_{x_{i,t+1} > x_{i,t}}$$
$$F_d(x_{i,t}, x_{i,t+1}) = \mathbf{1}_{x_{i,t+1} < x_{i,t}}.$$

$$\mathcal{G}_{EVGS} = \{(X_i, F) | X_i \in \mathcal{X}, F \in \mathcal{F}_{change}\}.$$

$$r(s_t, (X_i, F), a, s_{t+1}) = F(O(s_t)_i, O(s_{t+1})_i),$$

O is the mapping from state to variable values

$$a \in \{(X_j, F') | X_j \in X_{pa(i,C)}, F' \in \mathcal{F}_{change}\} \bigcup \mathcal{A}.$$

Methodology

Algorithm 3 SubgoalTraining

Parameter subgoal-based hierarchical policy π_h; causality graph C; Candidate controllable variables set S_{CC}; Verification threshold ϕ_{causal}

Change Function Set $F = \{f_i, f_d\}$
Candidate goals $G_C = \{(X, y) | X \in S_{CC}, y \in F\}$
$k = MaxDepth(C)$
if $k > \pi_h.levels$ **then**
    $BuildNewLevel(k, \pi_h)$
end if
$InsertSubgoal(G_C, \pi_h)$
$k_{min} = MinDepth(G_C)$
$k_{max} = MaxDepth(G_C)$
for $k_{idx} = k_{min}$ **to** k_{max} **do**
    Trainning goals $G_T = \{g | g.depth = k_{idx}\}$
    Trained steps $t = 0$
    while $t < T$ **do**
        Random goal $g = RandomSelect(G_T)$
        Execute g and put $trajectory$ into replay buffer D
        Train π_h using D
        $t = t + Length(trajectory)$
    end while
end for
Controllable variables set $S_C = \{X | SuccessRatio((X, y)) > \phi_{causal}, y \in F\}$
return S_C

- check its depth first to decide whether to build a new subgoal level.
- insert the new subgoals to the corresponding levels.
- ensure the agent will not forget the old subgoals when training new subgoals
- return variables whose corresponding subgoal success ratio exceeds the pre-defined ratio ϕcausal as controllable variables.

CDHRL Structure

Algorithm 1 CDHRL

Parameter Threshold ϕ_{causal}

initial SCM's structure parameters η, functional parameters θ; subgoal-based policies π_h
Initial the intervention variables set $S_{IV} = \{V_{action}\}$
while $True$ **do**
  Intervention data $D_I = InterventionSampling(\pi_h, S_{IV})$
  Causality graph $C = CausalityDiscovery(\eta, \theta, D_I, S_{IV})$
  Candidate controllable variables set $S_{CC} = \{V_i \mid V_i \notin S_{IV} \text{ and } V_{pa(i,C)} \subset S_{IV}\}$
  if S_{CC} is empty **then**
    Break
  else
    Controllable variable set $S_C = SubgoalTraining(\pi_h, C, S_{CC}, \phi_{causal})$
    $S_{IV} = S_{IV} + S_C$
  end if
end while
initial upper policy π_t over subgoal-based policies π_h
train π_t maximizing the calculated extrinsic reward

- first select new effect variables, whose cause variables have been in the controllable intervention variables set S_{IV} , as candidate controllable variables S_{CC} .
- train subgoals of the candidate controllable variables.
- the subgoal training success ratios are compared with the pre-defined threshold Φ_{causal} to select successfully trained subgoals.
- add new controllable variables S_C that with successfully trained subgoals to the intervention variables set SIV before the next round of intervention sampling and causality discovery.

Baselines

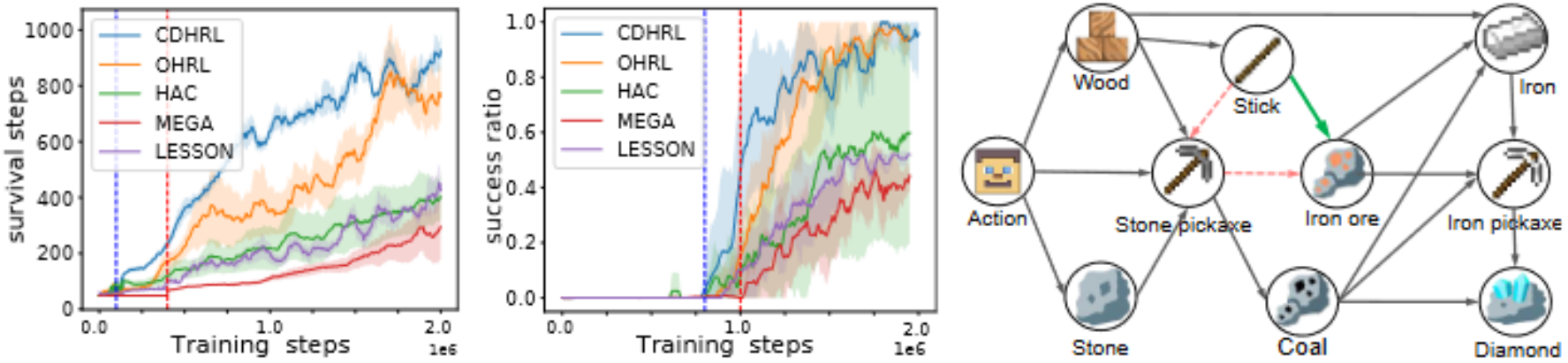
Oracle HRL (OHRL) is implemented as a two-level DQN with HER and an oracle goal space. The goal space is a subset of GEV GS after artificial eliminating unreachable and useless subgoals.

HAC is a powerful goal-conditioned HRL that discovers subgoals with a randomness-driven exploration paradigm. We implement a two-layer HAC with subgoal space GEV GS

LESSON is a modified goal-conditioned HRL method based on HAC that discovers subgoals from slowly changed features

MEGA is a kind of goal-conditioned HRL enhanced by curriculum learning, which pre-train subgoals in the order of their training progress. We set the initial subgoal distribution as the uniform distribution on GEV GS and pre-trains subgoals with enough steps

Experiments



(a) Eden (b) 2D-Minecraft (c) Causality Graph in Minecraft

Figure 2: (a) Agent's survival time in Eden. (b) Success ratio in 2D-Minecraft. The vertical dotted lines indicate the end of pre-training of CDHRL and MEGA. Results are derived from average data in 8 trials. (c) The causal graph of 2D-Minecraft discovered by the agent. Some uncontrollable variables that unlinked are ignored here.

Experiments

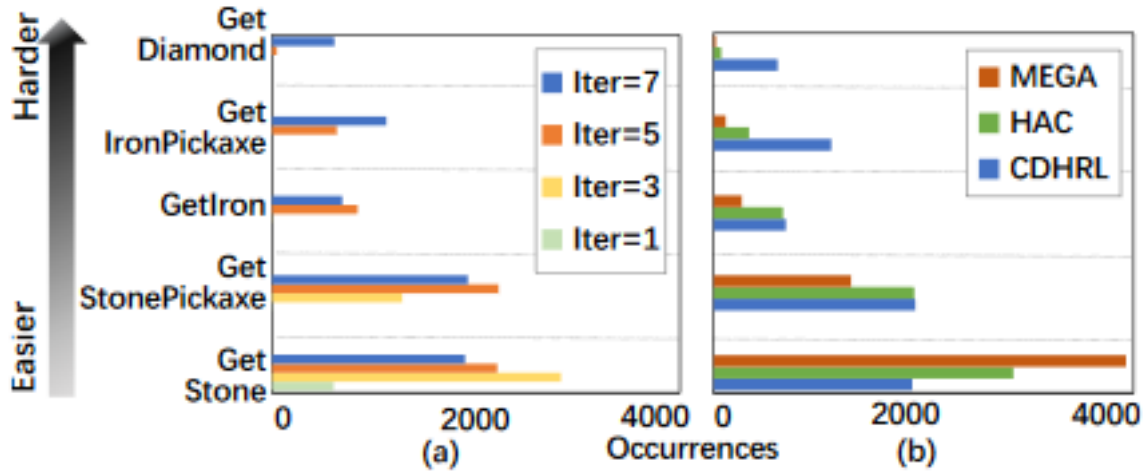
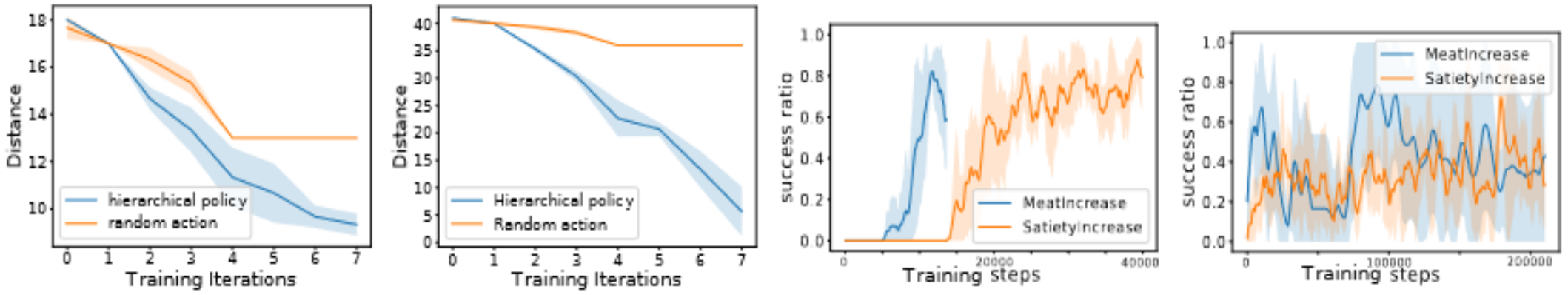


Figure 3: Exploration capability comparison. We select five explore milestones in 2D-Minecraft from easy-to-explore to hard-to-explore and record the occurrences in 10K test episodes to compare the agent's exploration capability. More occurrences on the hard-to-explore milestone represent higher exploration capability. (a) CDHRL's exploration capability iteratively increases along with the construction of hierarchical structures. (b) CDHRL shows much better exploration capability than HAC and MEGA. All compared methods are tested after trained 800K steps.

Experiments



(a) SHD (b) SID (c) CDHRL (d) MEGA

Figure 4: (a) Structure Hamming Distance (SHD) and (b) Structural Interventional Distance (SID) between learned and ground-truth causality graph. The intervention data sampled by the assistance of hierarchical policy make the causality graph more accurate. The ground-truth causality graphs of Eden and 2D-Minecraft are in Appendix C.2.(c) and (d) are learning curve of two subgoals.