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

Fast Counterfactual Inference for History-Based Reinforcement Learning

AAAI | 2023

Background

- **History-based Decision Process (HDP)**

  - ⟨O, H, A, T, P, R, γ⟩

    O： the observation space
    H： the history space
    A： the action space
    T： the horizon(time steps into the future when making decision)
    P: the history-action transition probability
    R: reward function
    γ: discount rate

Background

- **History-based RL**
  - current state + sequence of past states and actions = Decision Making
  - **Pros**
    - Partial Observability
    - Handling Sequential Dependencies
    - Adaptive Behavior
  - **Cons**
    - Data efficiency
    - Computational Complexity
    - **do not adjust for confounding correlations** → Counterfactual Inference (CI)

Background

- **Experiment**

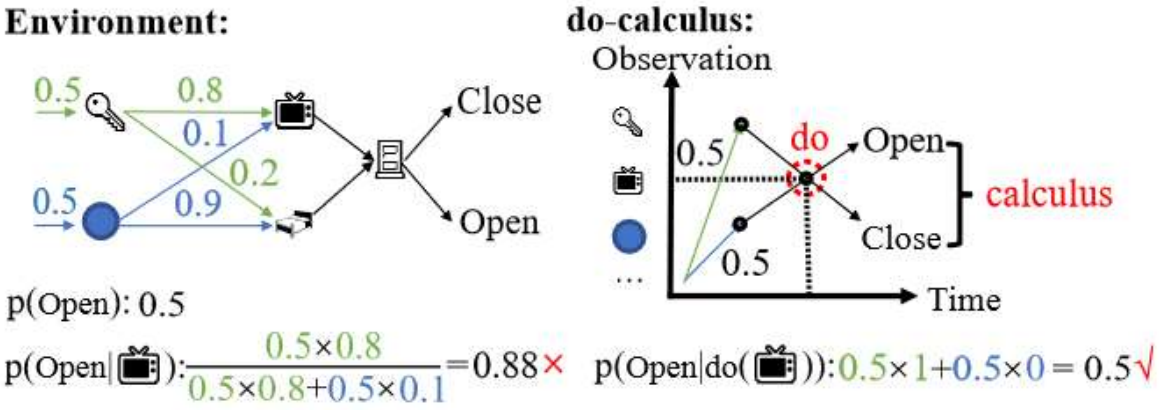


Figure 1: Key-to-door example. The high correlation on TV caused by sampling can be eliminated by do-calculus which separates confounders (key and ball).

Background

- **Former Effort**
  - Prior researchers perform CI on Markovian RL tasks for feature selection **where the states are Markovian**

单变量干预

  - for history-based RL, we should intervene for **each non-Markovian observation-and-time combination** ot and estimate its causal effect

不适用

high computational complexity due to the large scale of historical observations

需要开发多变量干预

Preliminaries

- **Do-calculus and Backdoor Formula**
  - **causal diagram G:**
    - covariates X := {X1, · · · , Xn}
    - a response variable Y
  - **the intervention operation do(Xj = x) :**

    adopted by imposing a certain value x on one of the covariates Xj

$$p\big(Y|\mathrm{do}(X_j = x)\big) = \int p\big(Y|X_j = x, \boldsymbol{X}^{\mathrm{ba}}\big)\mathrm{d}p(\boldsymbol{X}^{\mathrm{ba}}),$$

Xba ⊆ X are the backdoor variables relative to (Xj, Y)

1) Xba contains no descendant of Xj
2) Xba blocks each path pointed to X between Xj and Y

Preliminaries

- **Fine-Grained History Counterfactual Inference**
  - **Covariates**

    the past observations, the current observation, and the action {ht−1, ot, at}

  - **response variables**

    and the immediate reward and next observation {rt+1, ot+1}

Proposition 1 (Fine-grained CI). *Given* o_t *and* a_t, h_{j-1} *satisfies the backdoor formula relative to* $(o_j, \{o_{t+1}, r_{t+1}\})$ *for any historical observation* $o_j (j < t)$, *and the causal effect of* $\mathrm{do}(o_j = o)$ *can be estimated with*

$$p(o_{t+1}, r_{t+1} | \mathrm{do}(o_j = o), o_t, a_t) \tag{2}$$
$$= \int_{h_{j-1} \in \mathcal{H}_{j-1}} p(o_{t+1}, r_{t+1} | o_j = o, h_{j-1}, o_t, a_t) \mathrm{d}p(h_{j-1}).$$

Preliminaries

$$p\left(o_{t+1}, r_{t+1} | \mathrm{do}(o_j = \mathbf{0}), o_t, a_t\right) \tag{3}$$
$$= \int_{h_{j-1} \in \mathcal{H}_{j-1}} p(o_{t+1}, r_{t+1} | o_j = \mathbf{0}, h_{j-1}, o_t, a_t) \mathrm{d}p(h_{j-1}).$$

- **Average Treatment Effect (ATE)**

The difference between the counterfactual and factual distributions

$$\mathrm{ATE}(o_j^{\overline{o,o'}}) := \sum_{t=j+1}^{T} \mathbb{E}_{\substack{o_j \in \{o, o'\}, \\ o_t \in \mathcal{O}, a_t \in \mathcal{A}}} \tag{4}$$
$$\left|\left|p(\cdot | o_j, o_t, a_t) - p\left(\cdot | \mathrm{do}(o_j = \mathbf{0}), o_t, a_t\right)\right|\right|_1.$$

ATE趋向于0表示无因果

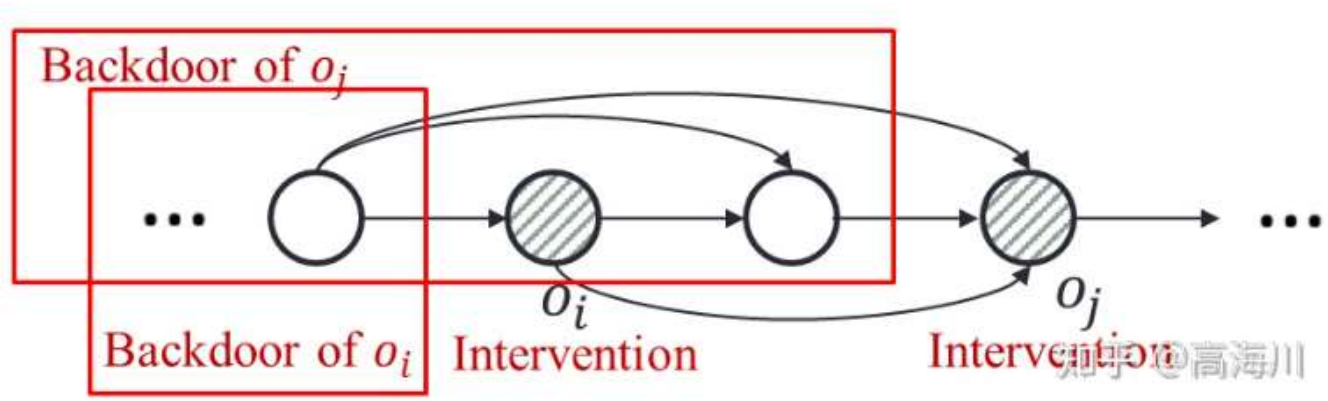
- History-base 需要遍历所有的历史观测，计算量大

Can we develop a method that can decide the causality amont numbers of variables?

Coarse-to-Fine History Counterfactual Inference

- **Steps**
  - estimate the causal effects of observations without timestamps. $\Omega(T \cdot |O|) \rightarrow \Omega(|O|)$. 空间推理
  - Second, perform CI on observation sub-spaces. $\Omega(|O|) \rightarrow \Omega(\log|O|)$ 时间推理

- **Problem**



multiple intervened historical observations have no common backdoor variables

Coarse-to-Fine History Counterfactual Inference

- **Counterfactual Inference on Observations**

Definition 1 (Step-backdoor adjustment formula). $\boldsymbol{X}_j^{\mathrm{s-ba}} = \boldsymbol{X}_j^{\mathrm{ba}} \setminus (\boldsymbol{X}_i^{\mathrm{ba}} \cup \{X_i\})$ *is step-backdoor relative to* (X_j, Y) *if 1)* $\boldsymbol{X}_j^{\mathrm{s-ba}}$ *has no descendant of* X_j*, 2)* $\boldsymbol{X}_i^{\mathrm{ba}}$*,* X_i*, and* $\boldsymbol{X}_j^{\mathrm{s-ba}}$ *block each path between* X_j *and* Y*, and 3) conditioned on* X_i *and* $\boldsymbol{X}_i^{\mathrm{ba}}$*, the distribution of* $\boldsymbol{X}_j^{\mathrm{s-ba}}$ *is identifiable.*

Theorem 1. *Given a set of intervened variables with different timestamps, if every two temporally adjacent variables meet the step-backdoor adjustment formula, then the overall causal effect can be estimated with*

Coarse-to-Fine History Counterfactual Inference

$$
\begin{aligned}
&p(Y|\mathrm{do}(X_i = x, X_j = x', X_k = x'', \cdots)) \\
&= \int \cdots \int p(Y|\boldsymbol{X}_i^{\mathrm{ba}}, \boldsymbol{X}_j^{\mathrm{s-ba}}, \boldsymbol{X}_k^{\mathrm{s-ba}}, \cdots, \\
&\qquad X_i = x, X_j = x', X_k = x'', \cdots) \\
&\quad \mathrm{d}p(\boldsymbol{X}_i^{\mathrm{ba}}) \\
&\quad \mathrm{d}p(\boldsymbol{X}_j^{s-\mathrm{ba}}|X_i = x, \boldsymbol{X}_i^{\mathrm{ba}}) \\
&\quad \mathrm{d}p(\boldsymbol{X}_k^{s-\mathrm{ba}}|X_i = x, \boldsymbol{X}_i^{\mathrm{ba}}, X_j = x', \boldsymbol{X}_j^{s-\mathrm{ba}}) \cdots
\end{aligned} \tag{6}
$$

计算Xi后门（后门准则）
计算Xj和Xi的step-backdoor（SBAF）

Theorem 2 (CI on observations). *Given o_t and a_t, the causal effect of* $\mathrm{Do}(o)$ *can be estimated by*

$$
\begin{aligned}
&p(o_{t+1}, r_{t+1}|\mathrm{Do}(o), o_t, a_t) \\
&= \int_{h_{t-1} \in \mathcal{H}_{t-1}^o} p(o_{t+1}, r_{t+1}|h_{t-1}, o_t, a_t) \mathrm{d}p(h_{t-1}|\mathcal{H}_{t-1}^o),
\end{aligned} \tag{7}
$$

where $\mathcal{H}_{t-1}^o$ denotes the history sub-space where each history $h_{t-1} \in \mathcal{H}_{t-1}^o$ contains at least one observation with value o.

$\Omega(\mathsf{T} \cdot |\mathsf{O}|) \rightarrow \Omega(|\mathsf{O}|)$.　空间推理

Coarse-to-Fine History Counterfactual Inference

- **Counterfactual Inference on Sub-Spaces**

Proposition 2 (CI on observation sub-spaces). *Given o_t and a_t, the causal effect of $\mathrm{Do}(\mathcal{O}^i)$ can be estimated by*

$$p(o_{t+1}, r_{t+1} | \mathrm{Do}(\mathcal{O}^i), o_t, a_t) \quad (8)$$
$$= \int_{h_{t-1} \in \mathcal{H}_{t-1}^{\mathcal{O}^i}} p(o_{t+1}, r_{t+1} | h_{t-1}, o_t, a_t) \mathrm{d}p(h_{t-1} | \mathcal{H}_{t-1}^{\mathcal{O}^i}),$$

where $\mathcal{H}_{t-1}^{\mathcal{O}^i}$ represents the history subspace where each history $h_{t-1} \in \mathcal{H}_{t-1}^{\mathcal{O}^i}$ contains at least one observation belonging to the observation subspace $\mathcal{O}^i$.

Proposition 3 (Coarse-to-fine CI). *If $Z \geq |\mathcal{O}^{\mathrm{Ca}}|$, the number of interventions for coarse-to-fine CI is $\Omega(\log|\mathcal{O}|)$.*

Ω(|O|) → Ω(log|O|)　时间推理

O^{Ca} : observation sub-space with historical causality.

Suppose:

an observation space is divided into Z ≥ | O^{Ca} | sub-spaces

at least Z − | O^{Ca} | parts containing no causal observation

Coarse-to-Fine History Counterfactual Inference

- **Combining RL and Coarse-to-Fine Counterfactual Inference**

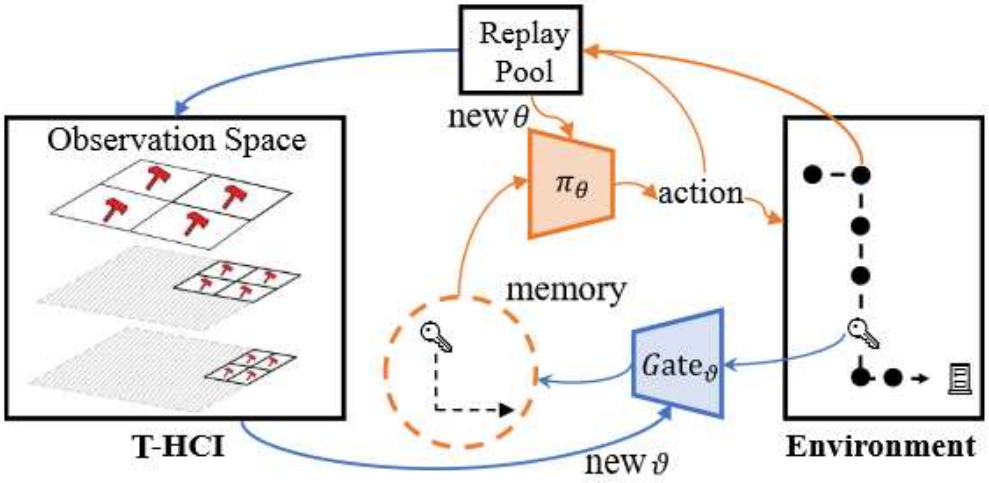


Figure 3: T-HCI Algorithm framework. The blue and orange lines respectively mark the CI loop and RL loop.

算法包含两个loops
1）T-HCI loop，
2）策略学习loop

两者交换进行：在策略学习loop里，agent被采样学习一定回合数量，并将样本存在replay pool中；在T-HCI loop中，利用存储的样本进行上述的因果推理过程

Coarse-to-Fine History Counterfactual Inference

- **Combining RL and Coarse-to-Fine Counterfactual Inference**

$\widehat{\mathcal{O}}$: the constructed discrete observation space

$\mathcal{O}^i$: the current intervened observation subspace

$\mathcal{O}^{\mathrm{w}}$: the observations that have been eliminated by previous inference

$$\mathrm{ATE}(\overline{\mathcal{O}^i, \mathcal{O}^{\mathrm{w}}}) \tag{9}$$
$$= \min_{\zeta} \sum_{t=1}^{T} \mathbb{E}_{\substack{h_t \in \mathcal{H}_t \\ a_t \in \mathcal{A}}} \left|\left| p_\zeta(\cdot | h_{t-1}^{\overline{\mathcal{O}^i, \mathcal{O}^{\mathrm{w}}}}, o_t, a_t) - \hat{p}(\cdot | h_t, a_t) \right|\right|_1.$$

I(·): the indicator function

$$\mathcal{L}_{\mathrm{Gate}}(\vartheta) = \frac{1}{|\widehat{\mathcal{O}}|} \sum_{o \in \widehat{\mathcal{O}}} |\mathrm{Gate}_\vartheta(o) - I(o \in \mathcal{O}^{\mathrm{w}})|, \tag{10}$$

Gateθ(oj) taking value 1 if and 0 otherwise

$$\mathrm{Gate}_\vartheta(o_j) = \begin{cases} 1 & \text{oj} \in \mathcal{O}^{\mathrm{w}} \\ 0 & \text{otherwise} \end{cases}$$

Coarse-to-Fine History Counterfactual Inference

- **Policy learning**
  - Built on the memories of causal historical observations,
    mapping $\phi : \{o0, \cdots, ot-1, ot\} \mapsto \{Gate(o0) \ominus o0, \cdots, Gate(ot-1) \ominus ot-1, ot\}$
  - name ψ as causal memory. a policy $\pi\theta(at|\phi\theta(ht))$
  - Train with Deep Monte Carlo (DMC)， A2C and PPO

Coarse-to-Fine History Counterfactual Inference

Algorithm 1 PPO, Actor-Critic Style

for iteration=$1, 2, \ldots$ **do**
  for actor=$1, 2, \ldots, N$ **do**
    Run policy $\pi_{\theta_{\text{old}}}$ in environment for T timesteps
    Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
  end for
  Optimize surrogate L wrt θ, with K epochs and minibatch size $M \leq NT$
  $\theta_{\text{old}} \leftarrow \theta$
end for

for iteration 1,2… do,
  for iteration 1,2,… do，
    PPO algorithm（update mapping ϕ ）
  end for
  store θ into replay pool
  for iteration 1,2…, do
    T-HCL algorithm（update ϑ）
  end for
end for

Experiments

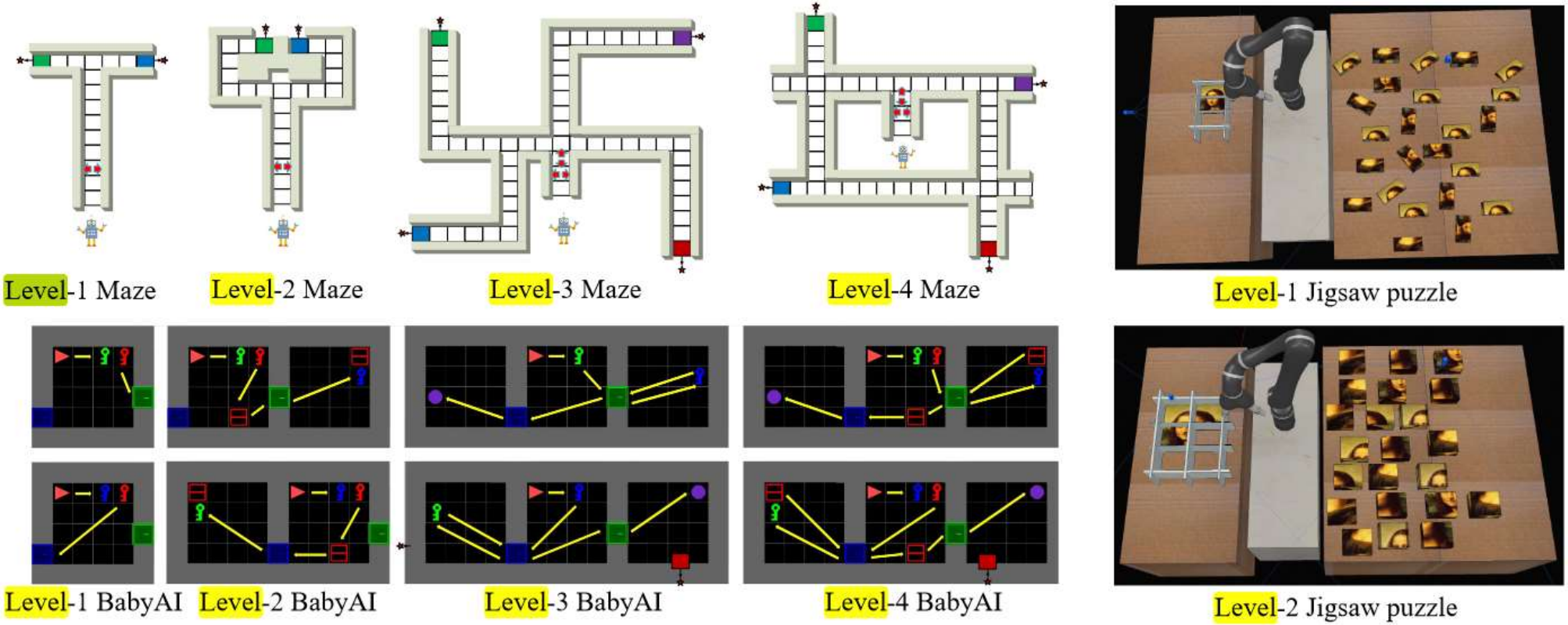


Figure 4: Environments of Maze, BabyAI, and Jigsaw Puzzle tasks.

Experiments

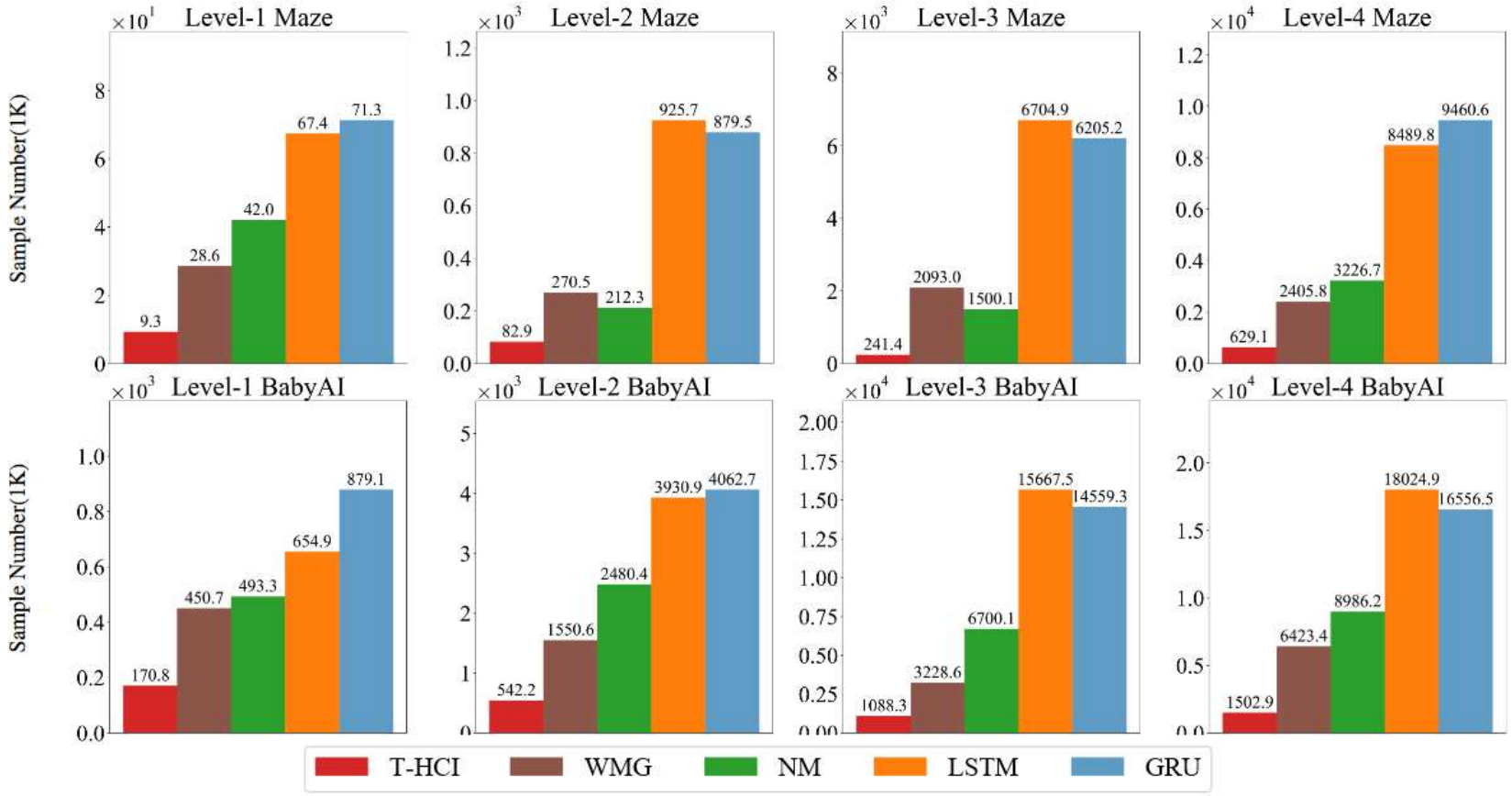


Figure 5: Comparison of average numbers of samples (thousand) in 10 trials of the Maze and BabyAI tasks.

Experiments

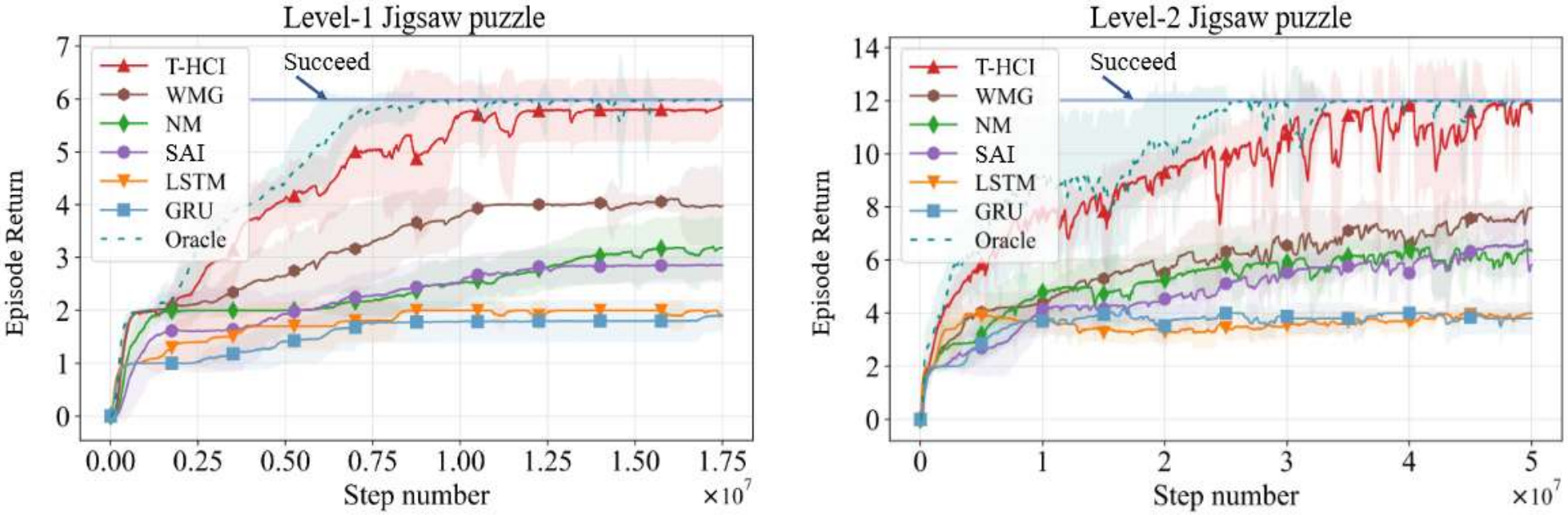


Figure 6: Learning curves of the Jigsaw puzzle tasks.