階層的世界モデルにおける現状と課題：Hierosの限界と将来の展望

Current Status and Challenges in Hierarchical World Models: Identifying Limitations of Hieros and Future Prospects

三好 理輝 *1　劉 智優 *2　山田 *3
Riki Miyoshi　Jiu You　Third Author's Name

*1ケンブリッジ大学　*2所属　*3所属
University of Cambridge　Affiliation #2 in English　Affiliation #3 in English

ここに
アブストラクトを
5 行程度
書き
ます

1. はじめに

2. 研究背景・目的

階層的強化学習 (HRL) は、探索において有益であると知られており [HRL benefit]、効率的な報酬割り当て、転移学習、解釈性などといった点で優位性があると考えられている [HRL survey]。

また、世界モデルは高いサンプル効率を実現し、エージェントが少ない環境との相互作用で学習することを可能にする。

この 2 つの要素を組み合わせたモデルに、Director[Director]、Hieros[Hieros] があり、ベースラインとして使用されている世界モデルを上回る性能を示している。一方で、解釈性やモデルをスケールさせる方向での研究は知られておらず、階層的な世界モデルの理論および評価実験は知られていない。

3. 関連研究

3.1 世界モデル

世界モデルは、「説明」。世界モデルは大きく、1. 状態空間モデル、2. 状態予測モデル、3. 観測予測モデル に分けることができる。代表的な世界モデルとして、Dreamer および TD-MPC2 が挙げられる。Dreamer は、状態空間モデルの一つであり、

3.2 階層的強化学習 (HRL)

3.3 階層的世界モデル (HWM)

階層的世界モデルも、世界モデルと同様、状態空間モデルである Dreamer の方策を階層化したものが Director である。マネージャーとワーカーの 2 つの方策が存在し、マネージャーはより長い時間軸で目標とする状態を圧縮された形でワーカーに渡し、ワーカーは通常の強化学習と同様の短い時間軸で行動を出力する。ワーカーの損失関数にマネージャーの提示した目標がどの程度達成されたかを組み込むことで、ワーカーとマネージャーの 2 つの方策がそれぞれ役割分担しながら同時に学習できるようにしている。

Director の拡張として Hieros がある。Hieros では、

1. 方策だけでなく世界モデルも階層化している。
2. 2 階層だけでなくそれより多い階層でも実験を行っている。
3. 計算効率を上げるために、状態空間モデルとして RSSM ではなく S5WM を用いている。
4. リプレイバッファからのサンプリングを工夫し、偏りを減らすことと時間計算量の削減を実現している。

という工夫が加えられたモデルが提案されており、Atari において DreamerV3 を超える性能を示した。本稿では方策だけでなく世界モデルも階層化しているという点で Hieros に着目し、その限界と可能性を調べる。

状態予測モデルである TD-MPC2 を階層的にしたものが Puppeteer である。これは、「説明」

4. 実験・考察

本稿では、階層的世界モデルの Hieros[Hieros] モデルの評価を行った。特に、論文中では検証されていなかった、長期タスクの評価に適している Visual Pinpad 環境での検証を行った。また、Atari でのモデルの内部状態の確認を行った。

5. 手法

Hieros のモデル検証では、既存のレポジトリ [Hieros Repo] を変更して実験を行った。また、Director ではこのレポジトリ [Director Repo] を用いて実験を行った。RTX-5070Ti または RTX-6000 一つを用いて学習をした。

5.1 Pinpad

Pinpad における Hieros の学習状況を調査した。まずは、Hieros のベースラインでの検証を行った。

1. 報酬設計の変更
2. 報酬割り当て係数の変更 Hieros では、各レイヤごとの方策は、external reward, subgoal reward と intrinsic reward の 3 つを用いて学習を行っている。
3. 方策エントロピーの変更

5.2 Subgoal 可視化

より詳しく Hieros の性能について調査するために、Hieros の各レイヤが提示している状態の可視化を行った。元の論文では過去の経験を参考にしたノイズを加えての可視化を行ってい

連絡先: 氏名，所属，住所，rm2278@cam.ac.uk

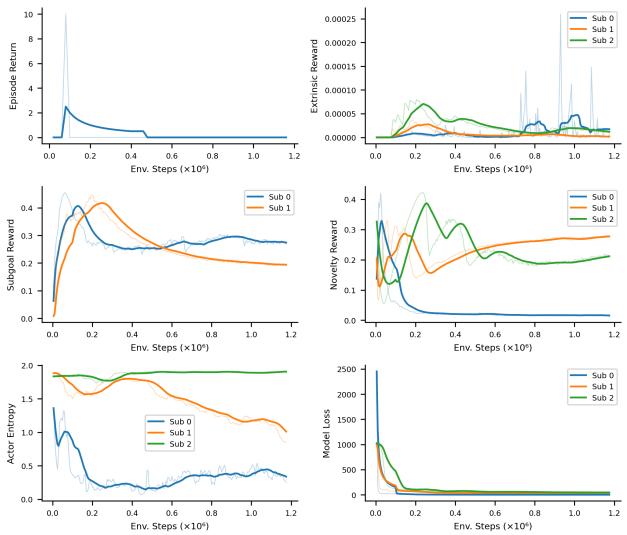

図 1: Visual Pinpad 環境における Hieros の学習および内部報酬曲線。

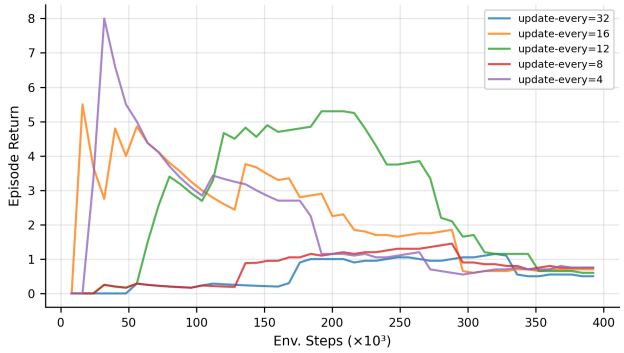

図 2: 異なる subactor-update-every パラメータでの学習曲線の比較。

たが、モデルの予測のみに焦点を当てるために、ノイズなしでの可視化を行った。

5.3 探索位置のヒートマップ

5.4 結果

1. 時間的抽象化方向

update-every は、それぞれの階層が、下の階層と比較してどの程度時間的抽象化を行っているかを表している。例えば、update-every=4 では、4 ステップごとに上の階層が下の階層に与えるサブゴールを更新しているという意味である。図からわかるように、多少の差異がある一方で、最終的に報酬が 1 になる程度に方策が収束している。

5.5 考察

なぜ方策がベースラインコードにエラーが含まれている可能性は否めない一方で、世界モデルを階層的にすることによって学習が不安定になっている可能性も高い。特に世界モデルは目的関数や誤差蓄積に起因して、また階層的モデルは階層の同時学習による非定常性によってハイパーパラメータに繊細であることが知られており、それらの組み合わせによって学習が不安定になってしまっている可能性もある。この理論的、実証実験的解析は今後の課題である。

6. まとめ

7. 展望

本稿では Hieros の 3 つの報酬係数を手動で設定し比較を行ったが、HarmonyDream [HarmonyDream] のように、報酬を自動でバランスすると、より多様なタスクにおいて学習ができるようになると予想する。

8. Limitation

実験が一つの seed 値でしか行われなかったこと。これはモデル改善の方向に力を注いでいたためである。

参考文献

[Director] Hafner, D. et al.: Deep Hierarchical Planning from Pixels, NeurIPS (2022).

[Hieros] Mattes, et al.: Hieros - Hierarchical Imagination on Structured State Space Sequence World Models (2023).

[DreamerV1]

[DreamerV2]

[DreamerV3]

[TD-MPC]

[TD-MPC2]

[Puppeteer]

[HRL benefit] Nachum, O. et al.: Why Does Hierarchy (Sometimes) Work So Well in Reinforcement Learning? (2019).

[THICK]

[Hieros Repo] https://github.com/Snagnar/Hieros

[Director Repo] https://github.com/danijar/director

[HarmonyDream] Ma, H. et al.: HarmonyDream: Task Harmonization Inside World Models (2024).

[HRL survey] Pateria, S. et al.: Hierarchical Reinforcement Learning: A Comprehensive Survey, preprint (2021).

[1]

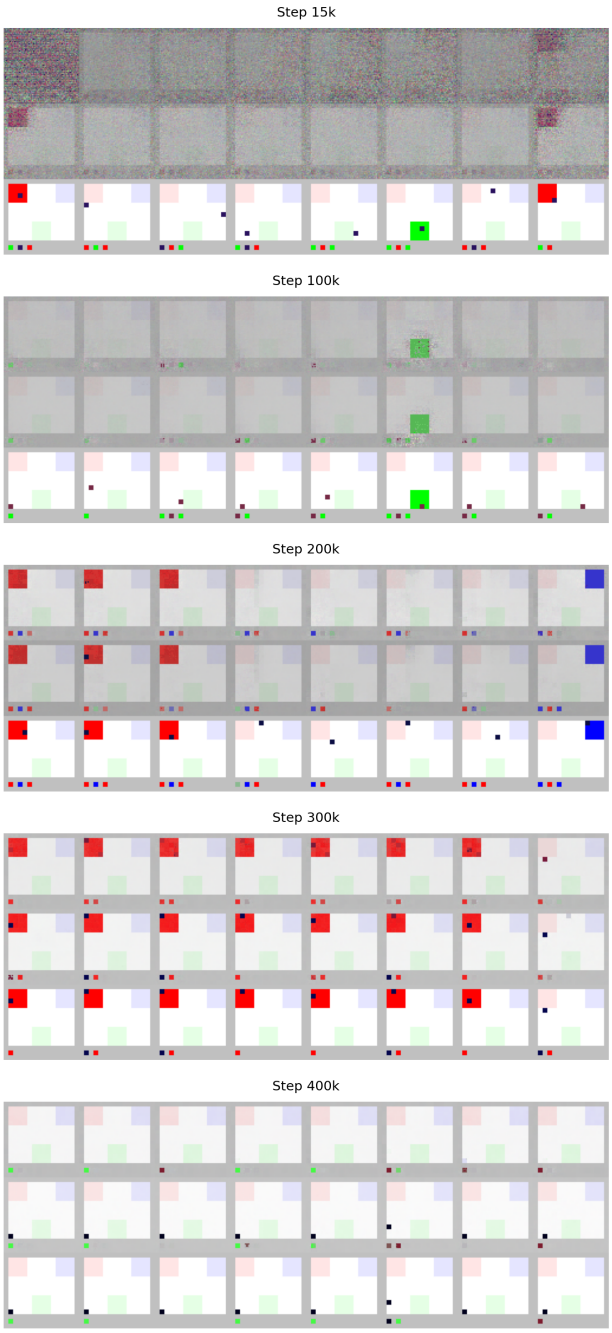

図 3: 学習過程におけるサブゴール可視化の時間的変化。

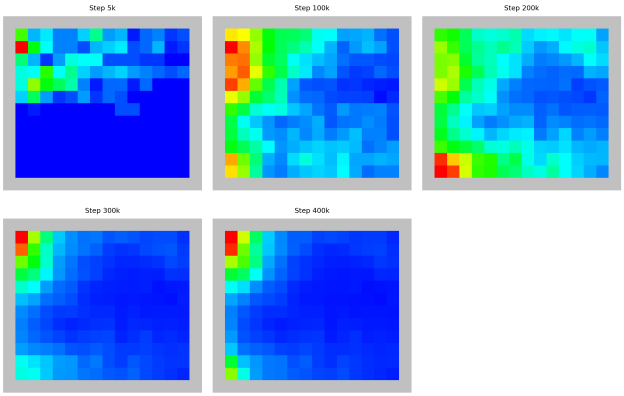

図 4: 探索位置のヒートマップの時間的変化。