世界モデルと言語の統合
~マルチモーダル基盤モデル+Mamba~
世界モデルチーム Kazuyuki

世界モデル(WORLD MODELS)とは何か

1. 人間は、限られた感覚で知覚できるものに基づいて、世界のメンタルモデルを開発する。私たちが下す決定と行動は、この内部モデルに基づいている。
2. 私たちの日常生活を流れる膨大な量の情報を処理するために、私たちの脳はこの情報の空間的側面と時間的側面の両方を抽象的に表現する。
3. ある瞬間に私たちが知覚するものは、脳が内部モデルに基づいて未来を予測することによって支配されていることも示唆されている

GENRL

引用元:[22] PIETRO MAZZAGLIA ET AL(2024)"GENRL:MULTIMODAL-FOUNDATION WORLD MODELS FOR GENERALIZATION IN EMBODIED AGENTS"
HTTPS://ARXIV.ORG/ABS/2406.18043

- マルチモーダル基盤世界モデル
- 現在の基盤であるビジョン言語モデル(VLM)は、ドメインギャップが大きいため、一般的に、具現化されたコンテキストで採用するために微調整やその他の適応が必要
- 本研究では、言語アノテーションを一切使用せずに、基礎VLMの表現をRLの生成世界モデルの潜在空間と接続し、整列させることができるマルチモーダル基盤世界モデルを提示することで、これらの問題を克服する
- 主な利点として、GenRLは言語アノテーションを必要とせず、事前訓練済みVLMを活用することで、言語データが不足するドメインでも適用可能

なぜ世界モデルを扱うのか

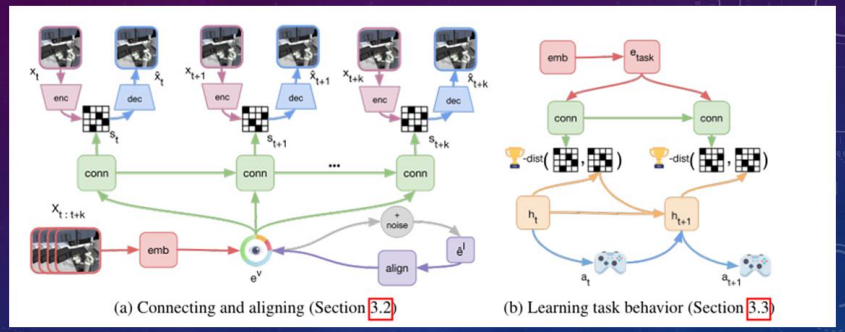
エージェントが環境との相互作用を通じて効率的に学習し、将来の状況を予測する能力を向上させるため

これにより、エージェントは試行錯誤を減らし、より迅速かつ効果的にタスクを遂行できるようになる

APPENDIX:LLMATCH世界モデルチームでの研究

GenRLの改善

- GenRLの問題点として、例えば複雑なテーマになると再構成画像の画質が悪くなることがあげられる
- この改善方法として、Mamba2を組み込んだりGRUをMamba2に置き換えたりすることでパラメータ数をTransformerほど上げることなく長期依存性が解消され、画質が向上することが見込まれる
- この改善に取り組んでいる



(a) Connecting and aligning (Section 3.2) (b) Learning task behavior (Section 3.3)