受験番号	※記入不要
氏　名	吉川　和之
現在の専門	電気電子情報工学
希望研究室	計算神経科学/BRI

小　論　文
（情報科学区分）

取り組みたい研究テーマ：Offline-MB-ERIL の Mamba 導入によるさらなる汎化性向上

１．これまでの修学内容

1.0　研究経緯

東京大学松尾研究室の講義を多数修了し、講義を通じて今後の人工知能の発展には脳科学の観点からの理解が必要だという強い探求心を持った。そこで世界モデルを使ったマルチモーダル基盤モデルの可能性を探るため、学部 3 年から東京大学松尾研究室 LLMATCH プログラムにて LLMCommunity 世界モデルチーム研究員として研究している。

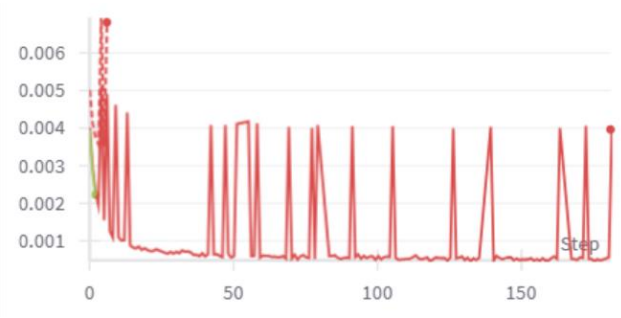

図 1: aligner_cosine_distance

1.1 課題

今まで様々なモデルが考えられているが、未だロボットシミュレーションや医療用マルチモーダル処理など、様々なアプリケーションに応用できるような普遍的な基盤モデルは生まれていない。

1.2 研究目的

私が行ってきた研究の主目的は、世界モデルを活用し、特にロボット応用を指向したマルチモーダル基盤モデルの精度を向上させることである。

そうしたモデルをロボットに搭載することで、ロボットの学習効率を上昇できることが想定される。

1.3 先行研究

私たちの日常生活に溢れる膨大な量の情報を処理するために、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。世界モデルはこれを深層学習で再現する技術である。

また近年、ロボットシミュレーション分野でマルチモーダル基盤モデルが注目されている。

Mazzaglia ら(2024)[2]は、VLM から視覚 or 言語を入力し、視覚の潜在表現と言語の潜在表現を統合することで世界モデルに強化学習を適用させた。このモデルを GenRL という。具体的には、事前学習済み VLM と世界モデルを繋ぎ、言語の潜在表現を視覚の潜在表現に直す。このモデルはマルチモーダル基盤モデルとして利用できる可能性もあると示唆されており、これによってロボットシミュレーションの精度が上がることが期待される。

1.4 課題と解決方法

先行研究は世界モデル上で入力画像が複雑になると入力の潜在表現を再構成した後の画質が悪くなる点に問題がある。これは時系列処理が上手くいっていないことが理由である。

Mazzaglia ら(2024)[1]によって用いられていた UnetDenoiser(ランダムにノイズを加えて、それをちゃんと復元できるように訓練することでモデルの頑健性を高めるためのもの)や世界モデルの RNN に対して Mamba(RNN よりも時系列処理に優れているモデル)を利用することにより、Transformer ほどのパラメータ数を必要とせず、aligner_cosine_distance(低いほど視覚の潜在表現と言語の潜在表現が統合しているという性能指標)を低くすることが可能かを試みている。

現在、この実験結果について論文を執筆中であり 9 月に学会発表を行う予定である。小規模実験の途中結果を図 1 に示す。これは UnetDenoiser に Mamba を加えた結果である。赤色点線が先行研究の結果、赤色実線が先行研究に Mamba を加えた結果である。Mamba を加えたことで aligner_cosine_distance が早く収束していることがわかる。

２．ATR で取り組みたい研究内容

2.0 志望理由・考察

Morimoto ら(2015)[2]によると、模倣学習は強化学習の中でも神経科学[3]に触発された分野であり、脳動機型ロボット工学で最も人気のあるアプローチの 1 つである。私は模倣学習のような脳の情報処理に類似したアルゴリズムを構築するアプローチに大きな興味を持った。また、BRI 研究室のように脳の仕組みを作りながら理解するというアプローチは、リハビリテーションというロボットと人間の両方に関わる分野にとって非常に有効なアプローチであると考えた。そのため、私は計算神経科学研究室で強化学習の研究を行いたい。

2.1 先行研究

深層強化学習で学習に必要なデータ数は膨大である。その対策として学習データ数を減らすために密な報酬を用いることが考えられるが、目的の行動を最適とする報酬関数を設計するのは容易ではない。報酬の設計を回避する手段としてエキスパート

受験番号	※記入不要
氏　名	吉川　和之
現在の専門	電気電子情報工学
希望研究室	計算神経科学/BRI

小　論　文
（情報科学区分）

(人間の振る舞いデータ)からの成功行動を準備し、そこから方策を学ぶ模倣学習がある。

模倣学習の中の一種として、敵対的生成模倣学習がある。これはエキスパートと学習エージェントの方策の状態・行動分布の違いを識別器と呼ばれる二値分類器で判別し、識別器から計算される報酬をもとに強化学習をすることで、エキスパートの方策に近づけるように学習者の方策を改善するものである。

従来の模倣学習では環境とのインタラクションを繰り返して学習する手法が多く、リアルな環境での学習に時間・コストがかかる。そこでUchibe(2023)[4]は Offline-MB-ERIL(オフラインのモデルベース敵対的生成模倣学習)を提案した。

この手法の特徴は 3 つある。

1 つはオフライン（追加の環境実験なし）で模倣学習を行うことである。

2 つ目はモデルベース（環境のシミュレータを活用）とすることでデータ効率を高めることである。

3 つ目はエントロピー正則化である。エントロピーとは確率分布の「広がり（不確実性）」を表す量であり、エントロピー正則化では、学習の途中でこの「広がり」を大きく保つように制約を加える。具体的には、方策やモデルに不確実性を持たせる効果がある。図 2 に概要図を示す。

2.2 目的

本研究では、オフラインで安全にロボット技能を獲得する模倣学習フレームワークに時系列モデリング技術を統合することにより、より高精度に制御する方策を生成できる基盤技術を確立する。

2.3 課題と解決方法

従来の実装ではポリシーとモデルに通常の MLP（多層パーセプトロン）が用いられており、時間的な長期依存関係を捉えることが難しい。しかし、近年提案された Mamba は系列データモデリングにおいて MLP を凌駕する性能を示しており、Mamba を導入することによって Offline-MB-ERIL の更なる性能向上が期待される。

まず学習方策およびダイナミクスモデルにおける従来の MLP ブロックを Mamba Block に置き換える。具体的には Ota(2024)[5]を参考にし、状態 x および行動 u を結合した系列をトークン列とみなし、これに位置エンコーディングを加えた上で Mamba に入力する。トークン列から最終状態ベクトルをプーリングし、行動予測（方策）または次状態予測（環境モデル）を行う。

識別器（Dpolicy, Dmodel）については、Mamba ベースの潜在表現を用いて構築し、識別性能を向上させる。

訓練プロセスは Uchibe(2023)[5]を参考にし、PU 学習ベースの損失関数(陽性データと未ラベルデータから陰性データの情報を推定し、分類器を訓練する関数)を最小化する。

2.4 予想される結果

提案手法の評価についても Uchibe(2023)[5]を参考にし、双腕ロボット Nextage を用いた手先の運動制御の課題を用いる。従来の Offline-MB-ERIL と比較して、方策およびモデルの負の対数尤度を低減することが予想される。さらに、系列モデリング強化によるロバスト性向上により、異なる初期条件や障害物配置に対しても高い汎化性能が達成されることが期待される。

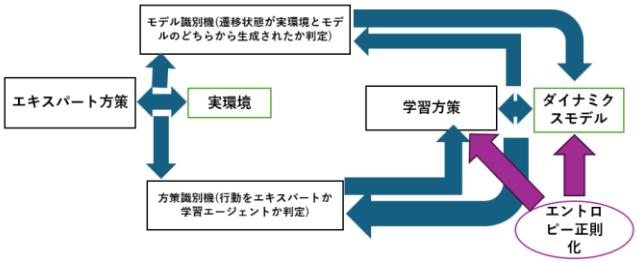

図 2 :Offline-MB-ERIL
(正確にはオフライン学習では実環境のデータが少ないため、エキスパートとは異なる別の方策を実環境に適用させたデータも使うことがある)

参考文献

[1] Pietro Mazzaglia et al(2024)"GenRL: Multimodal-foundation world models for generalization in embodied agents"

[2] Jun Morimoto et al(2015)"Creating the brain and interacting with the brain: an integrated approach to understanding the brain"

[3] Giacomo Rizzolatti et al(2004)"THE MIRROR-NEURON SYSTEM"

[4] Eiji Uchibe(2023)"Offline Model-Based Imitation Learning with Entropy Regularization of Model and Policy"

[5] Toshihiro Ota(2024)" Decision Mamba: Reinforcement Learning via Sequence Modeling with Selective State Spaces"