ボードゲームにおける強化学習エージェントの戦術分析に向けた可視化手法のプロトタイプ実装

丸中 翔太[1,a)]　伊藤 正彦[1,b)]

概要：ボードゲームプレイヤーは，強化学習エージェントから戦術を学ぶことが増えている．しかし，ボードゲームのエージェントには強い行動を選択するものが多く，ボードゲームプレイヤーの好む戦術に沿った提案をしていない．そのため，攻撃型や，守備型のような，それぞれの好む戦術を提案することのできるシステムが必要である．本研究では，報酬値を変更することで戦術を選択し，エージェントの行動を探索することができるシステムの可視化方法を提案する．このシステムでは，報酬値やハイパーパラメータを変更して学習をすることで，インタラクティブな可視化を実現する．ユーザーは，このシステムを使用することで，各々の好む戦術をエージェントから学ぶことができる．

1. はじめに

近年のボードゲームプレイヤーは，強い強化学習エージェントから学ぶことが増えている．ボードゲームプレイヤーには，攻撃型や，守備型のような，それぞれが好む戦術がある．しかし，強いエージェントは常に最適だと思われる行動を提示しているため，プレイヤーが好む戦術に沿った提案をすることは難しい．そのため，それぞれのボードゲームプレイヤーの戦術に合う行動の提案ができるシステムの開発が求められている．

また，強化学習には，学習率，割引率，報酬などの様々なパラメータがあるが，それぞれのパラメータを変化させた際に，エージェントの行動がどのように変化するのかを予測，探索するのは難しいという問題がある．パラメータの変化とエージェントの行動の変化を適切に探索できないと，エージェントがどのパラメータによって行動が変化したのかを理解することが難しくなる．そうすると，エージェントの行動を望んだものにするために，どのパラメータを変化させるのかを決定するのに時間がかかる．

そのため本研究では，パラメータを変化させることが容易であり，エージェントの行動の変化を探索しやすい可視化システムを開発することで，パラメータと行動の関係を探索する時間を短縮することができると考える．さらに，報酬にボードゲームの戦術の要素を加えることで，ボードゲームプレイヤーのそれぞれの戦術に沿った行動を提案できるシステムを開発することを目的とする．これにより，強化学習の学習者には，エージェントパラメータの関係性の探索に，ボードゲームプレイヤーには，それぞれが好む戦術に沿ったプレイングの上達に貢献するシステムを目指す．

[1] 北海道情報大学
[a)] s2221098@s.do-johodai.ac.jp
[b)] imash@do-johodai.ac.jp

2. 関連研究

McGregor[1] らは，マルコフ決定過程の開発におけるテストやデバッグを支援するための，汎用的なインタラクティブ可視化システム MDPVIS を提案した．MDPVIS は，報酬，モデル，方策のパラメータを対話的に変更し，出力の変化を可視化することができる．また，特定のドメインに依存しない設計であり，汎用的なインタフェースを用いて様々なシミュレータに接続できるようになった．これによって，モデルの妥当性の検証や，バグの特定を効率的に探索することができる．

Wang[2] らは，Deep Q-Network(DQN) のモデルを理解，診断，改善するための可視化システム DQNViz を開発した．DQNViz は，学習全体の統計的な推移から，特定のエピソードにおける行動を選択して分析することを可能にした．トレーニング後のモデルのパラメータを，統計ビュー，エポックビュー，軌跡ビュー，セグメントビューの 4 種類のカテゴリで可視化している．これによって，エージェントの行動や報酬パターンを深く分析することができる．また，Saldanha[3] らも，強化学習の理解を深める可視化分析ツールを開発した．エージェントの挙動を形で表現する squiggle を用いて直観的な比較を可能にした．

Mishra[4] らは，強化学習の知識がない人向けに理解を

支援するシステムを開発した．エージェントの行動をなぜ，なぜ行わなかった，いつの 3 種類の質問を行うことによって理解を支援した．これらの質問をテキストではなくフロー図で視覚化することにより理解しやすくなり，分析の負担を減らすことができた．

このように，強化学習エージェントを深く分析，探索する先行研究はいくつか存在するが，デバッグが目的のシステムが多い．また，パラメータを変更することで行動の変化を分析するシステムは少ない．さらに，ボードゲームの環境を題材として，ボードゲームプレイヤーの戦術に沿った提案をする研究は，著者らの知る限り存在しない．

3. 強化学習

本研究では，Q 学習を用いて開発を行う．Q 学習とは，エージェントの行動に対して適切な報酬を与えることで，エージェントの行動を強化する強化学習である．Q 学習は，強化学習の中でも一般的な手法であることに加えて，報酬の値や，いくつかのハイパーパラメータを変化させることでエージェントの行動が変化しやすいと考えた．

3.1 Q 学習

Q 学習では，Q 値をもとにエージェントの行動を決定している．学習率，割引率，報酬の値を使用して計算しており，それぞれの数値は，エージェントの学習速度，安定性，行動の特徴などに大きく関わっている．これらのパラメータのうち報酬の値を変化させることに重きを置いてシステムの開発を行う．報酬は行動の特徴に関わるパラメータであり，本研究の戦術を変化させることに最も適しているパラメータだと考えられる．

3.2 Q 学習の実装

Q 学習の実装方法として Gymnasium[5][*1]を使用する．Gymnasium とは，強化学習エージェントを訓練するための環境を提供する Python のオープンソースライブラリである．強化学習のシミュレーションが用意されており，エージェントが報酬を得て行動するという強化学習の基本的なループをサポートしているため，Q 学習を簡単に実装することができる．Gymnasium にはいくつかの実行環境が用意されているが，報酬を変化させるためにカスタム環境を利用する．カスタム環境は，ユーザー独自の強化学習タスクを定義することができる．これによって，既存の環境では実現できない独自の課題でエージェントを学習させることができるようになる．また，報酬の値を自由に変更することができるため，本研究に最も適していると考える．

[*1] 従来まで使われていた，OpenAI Gym のサポートが終了し，Farama Foundation が OpenAI Gym をコピーした後に独自のバージョンを追加したものを Gymnasium として提供している．

4. ボードゲームについて

4.1 ボードゲームの戦術

対戦型のボードゲームでは，プレイヤーによって戦術が分かれていることが多い．例えば，日本の麻雀プロプレイヤーには，一度のアガリでの点数を高くすることを好むプレイヤー，アガリまでの速度をできるだけ速くしたいプレイヤー，相手に払う点数をなるべく少なくしたい守備型のプレイヤーなどの様々な戦術のプレイヤーがいる．また，将棋にも中飛車や居飛車などの戦法があり，戦術があるボードゲームとして親しまれている．このように，ボードゲームには戦術が多く，ボードゲームプレイヤーは，好みの戦術が存在している．

近年ボードゲームプレイヤーが，AI を用いて強い戦術や行動を学ぶことが増えているが，提案される行動や戦術がプレイヤーの好みと離れていることがある．また，その行動が提案された理由が明確化されていないため，様々な場面での応用が利かないことが多い．そのため，プレイヤーの好みに沿って提案できる AI や，提案した理由を説明できる AI が必要である．そこで，本研究では，強化学習エージェントの戦術をユーザ自身でコントロールできるシステムの開発を目指す．

4.2 2 次元のランダムウォーク環境

本研究は，ボードゲームでの実装を目指している．しかし，ボードゲームの環境の実装は簡単ではないため，可視化手法の選定や提案手法のシステムの有用性を検証するためのテスト環境として，2 次元のランダムウォーク環境を選択した．ランダムウォークとは，次に動く場所が確率的に決まるすごろくのようなモデルである．このモデルは，計算量が少ないため，開発の初期段階におけるシステムの動作確認やデバッグに適していると考えられる．そのため，システムのデバッグや提案手法の有効性を検証したい本研究に最も適していると判断した．戦術の変化を探索するために，マイナス報酬である障害物マスを配置した．これによって，エージェントが障害物マスと同じマスに移動した際に設定したマイナス報酬を得ることとなる．

本研究では，エージェントが 1 回行動することを 1 ステップと呼び，エージェントがゴールするか，設定したステップ数エージェントが行動するまでを 1 エピソードと呼ぶ．これらと各報酬の値を調整することで，エージェントの行動に変化を与える．また，エージェントが 1 ステップで取ることのできる行動は，上下左右の 1 マスの移動である．左上をスタート地点として右下のゴールを目指して行動を行う．エージェントが盤面の外に移動しようとした際には，移動を行わずにそのまま 1 ステップの処理が終了する．報酬値には，1 ステップごとの報酬，動かなかった時

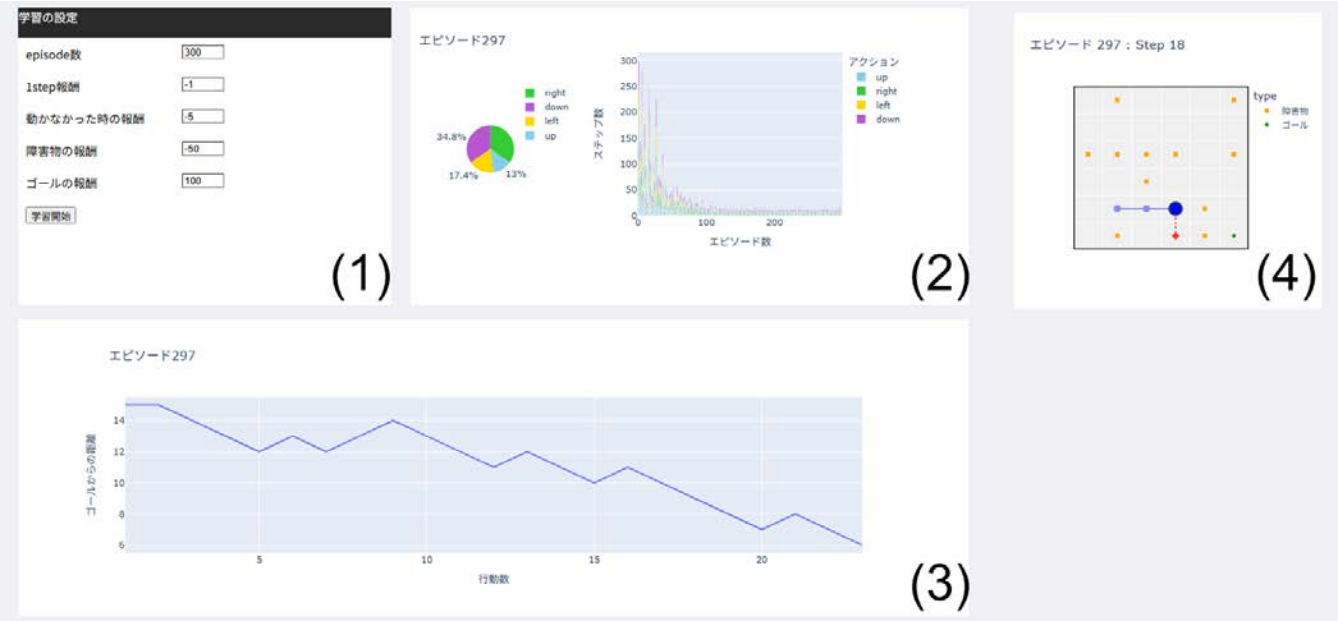

図 1　システムの全体図

の報酬，障害物の報酬，ゴールの報酬が含まれており，これらの値は変更することができる．

5.　可視化システムの実装

図 1 は，実際に開発した可視化システムである．パラメータ調整ビュー (1)，エピソード選択ビュー (2)，距離遷移ビュー (3)，ディスプレイビュー (4) の 4 種類で構成されており，インタラクティブな可視化を行っている．それぞれのビューから，入出力を行うことによって，ユーザーの求める情報を提供することができる．これらを使用することで，各エピソードでエージェントがゴールするまでの動きを詳しく探索することができる．

5.1　パラメータ調整ビュー (1)

パラメータ調整ビュー (1) では，エピソードの数，1 ステップごとの報酬，動けなかった場合の報酬，障害物の報酬，ゴールの報酬の 5 種類のパラメータを変更し，学習を開始することができる．それぞれに数値を入力した後に，ボタンを押すことによって，実際にエージェントが学習を行い他のビューに可視化を行う．数値の変更を行うことによって，ユーザーが重視しているものをエージェントの行動に反映することができる．

5.2　エピソード選択ビュー (2)

エピソード選択ビュー (2) では，パラメータ調整ビューによって入力した数値を元に学習した，各エピソードにおける行動回数の積み上げ棒グラフを表示する．このビューによって，各エピソードでのエージェントの行動量が俯瞰で探索できる．また，棒グラフをクリックすることで，エピソードを選択することができる．エピソードを選択することで，選択したエピソードの行動割合を示すパイチャートが表示される．さらに，選択したエピソードの詳細な分析をほかのビューで行えるようになる．表示されている色は，エージェントが行った各行動である．これにより，各エピソードでエージェントのどの行動が多かったかを簡単に分析することができる．

5.3　距離遷移ビュー (3)

距離遷移ビュー (3) では，エピソード選択ビューで選択されたエピソードのエージェントとゴールの距離を折れ線グラフで確認することができる．これによって，エージェントの動きを折れ線の形で簡単に予測することができる．さらに，折れ線グラフをクリックすることで，ステップを選択することができ，ディスプレイビューにてエージェントの詳細な動きを分析することができる．

5.4　ディスプレイビュー (4)

ディスプレイビュー (4) では，距離遷移ビューで選択されたステップから 2 ステップ前と 1 ステップ後のエージェントの位置を確認することができる．距離遷移ビューでは確認することのできなかったエージェントの詳細な位置が表示される．薄い青色の円は過去 2 ステップ，大きい青い円は選択したステップ，赤いダイヤで次のステップを表示している．これらを線でつなぐことでエージェントの動きを探索することができる．

6.　探索事例

システムの有用性を確認するために二つの探索を行った．

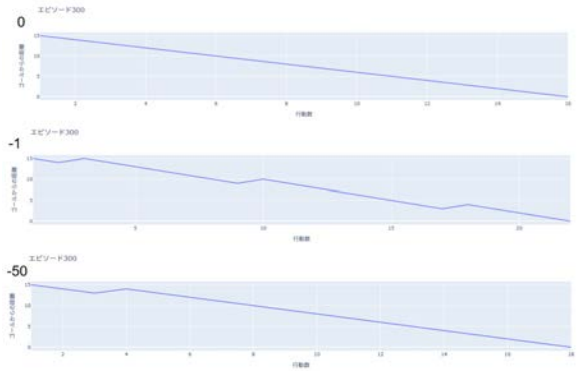

図 2 探索後の距離遷移ビューの比較

一つ目として，パラメータ調整ビューで値を変更することによって，変化するエージェントの動きを探索することができるかを探索した．

二つ目に，距離遷移ビューにて見られる特徴的な折れ線グラフを，ディスプレイビューで確認することによって，エージェントが実際にどのような動きをしているかを探索した．

6.1 報酬値を変更することによるエージェントの行動変化の探索

障害物の報酬の値を 0，-1，-50 に変更して学習し，最も学習が進んでいると考えられる 300 エピソード目を選択し，距離遷移ビューを探索した (図 2)．

障害物の報酬の値が 0 の場合，ゴールに直進していることが読み取れる．これは，障害物の報酬の値が 1 ステップ報酬と変わらないため，迂回の選択肢を取らないからだと考えられる．

報酬の値が-1 の場合は何度かゴールから離れていることから，障害物を交わしている可能性がある，ことが読み取れる．しかし，報酬の値が-50 の場合との差を読み取ることはできなかった．これは，-1 も-50 も受け取る必要のない報酬として一致していたからだと考えられる．

これらの探索により，報酬の値を変更することによって，エージェントの行動が変化し，それらを探索することができるシステムであることがわかった．

6.2 特徴的な折れ線グラフのエージェントの行動探索

様々な探索を行っている際に特徴的な折れ線グラフ (図 3) を複数発見した．それらの折れ線グラフが実際にどのような動きを行っているかの探索を行った．

折れ線グラフの形に注目をすると，ギザギザや横線が目立っていることがわかる．これらをディスプレイビューで確認することで，エージェントの実際の行動を探索した．

ギザギザの行動を探索するために，ギザギザの真ん中に位置するステップを選択した．そうすると，エージェントが二つのマス目を行き来していることがわかった．また，

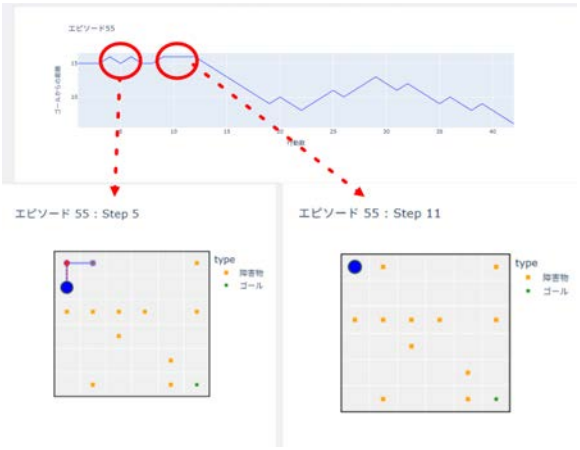

図 3 特徴的な折れ線グラフ

横線の際には，エージェントが壁に向かって進もうとしているために行動できていないことがわかった．

これにより，エージェントの実際の行動を詳細に探索することができた．さらに，他のエピソードの折れ線グラフを探索する際にも，これらの特徴が見られる場合には似た行動をとっていることが予測できるようになった．

7. まとめ

本研究では，利用者が自由にパラメータを変更し，その都度学習させることによって，エージェントの行動がどのように変化するかを探索できるシステムを開発した．また，6 章にてシステムを用いた探索をした結果，障害物の報酬の値を変化させることによって，行動が変化する場合があることが分かった．さらに，特徴的なグラフにおける実際のエージェントの行動を探索することができた．そのため，この手法は一定の有用性があると考えられる．

今後の課題として，ボードゲーム環境での実装をする必要がある．実装するボードゲームとしてガイスターを候補としているが，実際に実装可能かを探索する必要がある．さらに，ボードゲーム環境を適応する際に，Q 学習での処理は難しい可能性もある．そこで，本研究ではシステムを DQN を用いて開発することを考えている．また，本研究の評価手法の検討及び評価を行う．現時点ではユーザー評価を考えており，実際にシステムを使用してもらい，著者らで用意した質問に答えてもらうアンケート調査を行う予定である．

参考文献

[1] S.McGregor et al.: Interactive visualization for testing Markov Decision Processes: MDPVIS, *JVLC*, Vol. 39 (2017).
[2] J.Wang et al.: DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks, *TVCG*, Vol. 25, No. 1 (2019).
[3] E.Saldanha et al.: ReLVis: Visual Analytics for Situational Awareness During Reinforcement Learning Exper-

imentation, *EuroVis 2019 - Short Papers* (2019).
[4] A.Mishra et al.: Why? Why not? When? Visual Explanations of Agent Behaviour in Reinforcement Learning, *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)*, pp. 111–120 (2021).
[5] F.Foundation (2024): Gymnasium Documentation, `https://gymnasium.farama.org`, 2025-07-02 参照.