深層強化学習を用いたカードゲーム「デュエルマスターズプレイス」の自動対戦 AI
Deep Reinforcement Learning AI for DUEL MASTERS PLAY'S

川島　祥太
Shota Kawashima
法政大学情報科学部コンピュータ科学科
E-mail: shota.kawashima.8k@stu.hosei.ac.jp

Abstract

Recently, game AI with deep reinforcement learning (DRL) has shown high performance, and some AI has surpassed human players. There are studies on AI with DRL for strategy card games, e.g., Hearthstone, although such AI is not yet popular unlike that for shogi. This paper proposes AI with Deep Q-Network (DQN) for a strategy card game called DUEL MASTERS PLAY'S. This game has various strategy challenges such as resource management, creature card attacking, and card effects. The proposed AI introduces "one player's play" to learn the complex game efficiently. The method enables the AI to learn actions and rules that do not depend on the opponent by learning the rewards obtained by using cards in the environment without the opponent. This paper also describes the implementation of baseline AI, a deck selection model, and a DQN model for this game. This paper shows the effectiveness of "one player's play" by its total rewards and win rate when it learned and fought with the baseline AI in the environment with 38 cards and 5 decks. As a result of the experiment, the AI that learned by the proposed method used cards effectively in some degree and surpassed the baseline AI in the win rate.

1. はじめに

昨今では様々なジャンルのゲームを題材とし，深層強化学習によって学習する自動操作 AI の開発が行われている．それらの AI は高いパフォーマンスを見せるようになっており，対戦ゲームで人間のプロプレイヤーに勝利するレベルの AI も存在する．こうした研究はカードゲームの分野でも行われており，AI 研究用のカードゲーム Legends of Code and Magic が開発されており，それを題材に AI の性能を競うコンペティションが過去に開催された．また，コナミはデジタルカードゲーム「遊戯王マスターデュエル」を題材にした強化学習の AI を 2 つのエンジニアチームに分かれて開発し，AI 同士の対戦をイベントで披露した．しかし，強化学習を用いた高度なカードゲーム AI はユーザー向けのコンテンツで扱われることがまだ少なく，将棋 AI のように根付いてはいない．

本研究ではカードゲーム「デュエルマスターズプレイス」を題材に深層強化学習を用いた自動対戦 AI を開発する．そのために，複雑なカードゲームを学習する手法「一人回し」を提案する．一人回しは対戦相手を省いた学習環境により，ゲームの対戦相手に依存しない範囲を学習出来，効率的な機械学習を可能とする．その性能を示すために，一人回しの学習をカード使用の報酬値から評価し，本手法を導入した AI とそうでない AI をルールベース AI との対戦結果で比較した．その結果，一人回しで学習すると多少だが敵に依存しないゲーム環境を学べており，一人回しを学習に導入した AI は導入していない AI より優れていた．

2. 関連研究

本研究で用いる Deep Q-Network (DQN)は Mnih ら [1]が開発した手法であり，ニューラルネットワークを用いてゲーム画面の画像データを入力とした強化学習を行ったものである．題材のゲームは家庭用ゲーム機 ATARI2600 のゲームの内 7 つ（Beam Rider，Breakout，Enduro，Pong，Q*bert，Seaquest，Space Invaders）で，全てのゲームを同じネットワーク構造，学習アルゴリズム，ハイパーパラメーターで学習させた．結果として 6 つのゲームでは既存の手法より優れた成果を示し，3 つのゲームは熟練の人間プレイヤーを上回っていた．また Mnih ら [2]は DQN を用いた単一の AI モデルで ATARI2600 の 49 個のゲームを学習させた．結果 43 個のゲームで既存の手法より優れた結果を出し，29 個のゲームでプロのテスターのスコアの 75%以上のスコアを出した．

またカードゲームを題材とした研究について，Xi ら [3]が Legends of Code and Magic を題材として開発したデッキ構築と対戦の両方を行う AI がある．デッキ構築と対戦の 2 ステージを処理するための方策 end-to-end policy とカードゲームのような 2 プレイヤー制のゼロサムゲームを学習するための Optimistic Smooth Fictitious Play algorism の 2 つを導入し，結果として Legends of Code and Magic の AI の性能を競う Conference on Games (COG) 2022 の大会で優秀な成績を残した．また Xiao ら [4]は Legends of Code and Magic の研究を改良し，商業のカードゲームのハーストーンでも同様の AI を開発した．開発した AI は公式リーグ（中国地域）で優れた成績を残している人間プレイヤーに勝利するほどの性能であった．

Supervisor: Prof. Hiroshi Hosobe

3. 準備

3.1. デュエルマスターズプレイス

デュエルマスターズとは 2 プレイヤーでの対戦をメインとするカードゲームであり，デュエルマスターズプレイスとはデュエルマスターズのスマートフォン向けアプリケーション用のバージョンである（以後は単にデュエルマスターズと呼ぶ）．図 1 はその対戦画面である．

デュエルマスターズで基本的にプレイヤーはそれぞれ 40 枚のカードで構成される山札を使用して戦う．プレイヤーは試合中にマナと呼ばれるリソースを溜められ，マナを用いてカードを使用出来る．またプレイヤーはクリーチャー（使用すると場に出て，相手への攻撃などが可能なカードタイプ）や呪文（能力の発動のみを行うカードタイプ）などを駆使して試合を有利に進められる．クリーチャーカードで相手プレイヤーを攻撃すると相手のシールドを減らせ，シールドが無くなった相手に直接攻撃を行うことで対戦に勝利出来る．

デュエルマスターズでは，ターン開始時に手札を消費するかカードの効果でマナを増やせるが，手札からカードをマナゾーンに置く場合，手札から手放すカードを適切に選択する必要がある．またカードを使用する場合は，使用するカードと同じ文明（カードの属性）を持つカードをマナゾーンに置いている必要がある．そのため手札からカードをマナに置く行為はこのゲームの難しさの要因となっている．またクリーチャーカードの攻撃にも多くのリスクが含まれている．例えばクリーチャーが攻撃する場合，そのクリーチャーは相手からの攻撃が可能な状態となる（タップ状態）．クリーチャーが対戦相手に直接攻撃をする場合，攻撃したシールドは基本的に相手の手札に加わるため相手にリソースを与えることになる．その他にもクリーチャーの攻撃はリスクが存在するため，クリーチャーによる攻撃に関してもこのゲームの難しさの要因となっている．

図 1 デュエルマスターズプレイスの対戦画面

3.2. DQN

強化学習の 1 種である Q 学習は，状態に紐づいた行動の価値を示す Q 値を取り入れた学習であり，学習内での試行錯誤の過程で各状態における各行動の Q 値を修正していくことで AI の行動選択を適切なものにしている．Q 学習にニューラルネットワークを導入した深層強化学習 Deep Q-Network (DQN) [1] [2]では，状態を入力としたネットワークによって Q 値の出力をしている．通常の Q 学習で Q 値は行列で表現出来，行列の要素を 1 つずつ修正することで実装出来る．ただし，状態のとりうるパターンが膨大である場合，学習やプログラム上での表現が困難となる．しかし DQN を用いた場合，ネットワークのパラメーターを更新することで Q 値 1 つずつを直接修正するのではなく Q 値を出力するネットワークモデルを修正するため，膨大な数の状態パターンにも対処可能である．

4. 提案手法

本研究ではカードゲーム AI が対戦学習を行う前の「一人回し」の導入を提案する．一人回しとはカードゲームにおける練習行為であり，対戦相手がいない状況で自身のデッキを操作することである．対戦相手がいない状況であるため，対戦相手に干渉する行動に関して学習出来ないが，逆に対戦相手に依存しない行動やルールを優先して学習出来る．例えば本研究の対象であるデュエルマスターズではマナや手札などといったリソースの運用は比較的に対戦相手に依存せず，かつデュエルマスターズで重要な要素である．これを対戦相手との攻防の前に学習することで学習の質を向上する．このように本手法を導入することで複雑なカードゲームの学習をより効率的に行うことが期待出来る．

4.1. 提案手法の詳細

一人回しは対戦相手がいない状況でゲームを行うことであり，対戦相手に関係する行動を一切省いたものである．本研究では以下の流れで行う．

① ゲーム開始．
② プレイヤーのターン開始，山札の上からカードをドローする（最初のターンはランダムでカードのドローが出来ない）．またマナポイント＝マナに置かれているカード数とする．
③ 手札からマナゾーンにカードを置く（任意）．マナにカードを置く際は，マナポイントが 1 加算される．
④ マナポイントを消費してカードを使用する（任意）．
⑤ プレイヤーのターン終了，②に戻る．一定ターンが経った場合はゲームを終了する．

これは本来の対戦から自身の攻撃のフェイズと対戦相手のターンをなくしたものであり，ただカードを使い続けるものである．そのため本研究では上記の流れの④でカードを使用した際に次式の報酬を与える．

$$R = c \cdot (11 - t) \cdot v$$

ただし，cは使用したカードのコスト，tはターン数，vは補正値を示す．

この報酬はカードのコストに応じて大きくなるため，AI はコストが高く強力なカードを使用するようになる．さらにターンの経過に応じて小さくなる報酬でもあるため，AI はなるべく早めにカードを使用するようになる．これらにより AI はコストの高いカードを早く出すために手札やマナといったリソースを溜めることを覚えられる．また使用するデッキによってはリソースを溜めて強力なカードを出すことよりもコストが小さいカードを早期から出し続けることをメインとするものもある．報酬がタ

ーンの経過に応じて小さくなるという要素を持つため，そのようなデッキタイプを学習することも期待出来る．

DQN は基本的に通常の学習と同じように扱うが，通常の学習では対戦相手に関する状態を扱うのに対して一人回しは対戦相手に関する情報がない．そのため一人回しにおける対戦相手の状態は全て 0 で扱い，その入力が出力に影響を与えないようにする．一人回しの後に本来の学習が始まり，対戦相手の状態が正しく入力されることで対戦相手を加味した行動を学ぶようになる．またデュエルマスターズはカード効果の選択対象が存在しない場合その効果は不発となるため，対戦相手に作用する効果を持つカードを一人回しの環境で使用しても問題はない．

4.2. 提案手法を導入した学習

一人回しは対戦による学習の補助となるものであり，通常の学習の前に行う．そのため，一人回しを導入した学習は以下のようになる．

① 一人回しによる学習を行う．一人回しは 1 セットにつき 10 ターン行い，試行中はカード使用時に報酬が得られる．10 ターン経過後は経験データが十分あれば，集めた経験データを基にネットワークを更新する．その後ゲーム環境を初期化する．

② ①で学習したネットワークを用いて対戦による学習を行う．具体的な流れは 5.4.2 節で説明する．

5. 実装

5.1. デュエルマスターズの実装

本研究で実装したゲーム環境は元のゲームのルールやカードの効果を再現しており，ゲームの進行の面では元のゲームと概ね同様である（ただし AI の学習上一部変更している点がある）．また研究に使用するカードはデュエルマスターズプレイスの初期から使用可能なカードと第一弾のパックに封入されるカードの内から 38 種を用い，それによって出来る 5 デッキを使用可能なデッキとして設定した．デッキに関してはウェブサイト「【デュエプレ】全デッキリスト｜デッキコード一覧」[1]における「他過去環境デッキ整理中」の欄のものを採用した．図 2 は実装したゲーム環境の対戦画面である．

図 2 実装したゲーム環境の対戦画面

[1] https://kamigame.jp/デュエプレ/デッキ/index.html

5.2. ルールベース AI の実装

本研究でルールベース AI は対戦による学習や AI の性能評価に用いる AI であり，以下はルールベース AI のルールの一部である．

① ターン開始時にマナのカードを置く場合，マナにない文明のカードを優先的に置く．

② 呪文カードが使用出来る状況で，効果の対象が存在せず使用の意味がない場合はその呪文カードを使用しない．

③ クリーチャーカードの攻撃は自身の場にカードが 3 枚以上ある場合に行う．また自身の場のカードが 2 枚以下であっても，相手の場のパワーが最大のクリーチャー以上のパワーを持つクリーチャー（攻撃時のパワーを参照）は攻撃させる．

④ 自身のクリーチャーを攻撃させるとき，そのクリーチャーよりパワーが高いクリーチャーへ攻撃しないようにする．

⑤ 相手クリーチャーが攻撃する場合，そのクリーチャーよりパワーが低いクリーチャーでは攻撃のブロックをしない．ただし自身のシールドがない場合はパワーに関わらずブロックをする．

その他にもいくつかルールはあるが，以上のようにルールベース AI は極力無駄な行動を避けるように出来ているため，多少ランダム性はあるが完全にランダムな AI より遥かに高い性能を持っている．

5.3. デッキ選択モデル

デュエルマスターズのようなカードゲームでは対戦前にデッキを構築しておく必要がある．本研究ではデッキを予め 5 つ設定しそれらのみが使えるため，その中からデッキを選択する必要がある．デッキ選択モデルの構築のため，学習後に対戦を繰り返し行いその時の各デッキの勝利数を基にモデルを構築する．デッキの選択確率はソフトマックス関数で計算する．

5.4. DQN の実装

5.4.1. 状態・行動・報酬

AI がゲーム環境を認識しその中で行動を起こすには，AI が取得する状態と行える行動を定義する必要がある．入力と行動はそれぞれ表 1，表 2 で定義した．AI は行動を行うタイミングで，表 1 の状態をニューラルネットワークに入力し，表 2 の行動の Q 値を得る．AI は行動可能なものから Q 値が最も高い行動を選択する．なお選択可能な行動の情報は状態の入力時に同時に与えられ，行動を出力するタイミングで不可能な行動は制限される．また報酬は通常の対戦では勝利時に$+1$，敗北時に-1が与えられる．一人回しの学習では 4.1 節で説明した報酬とデッキのカードが尽きた際に-1の報酬が与えられる．

表 1 状態の定義

状態の種類	状態の内容
各ゾーンのカードの種類	自身の手札や場，相手の場などの確認可能なゾーンに存在するカードの種類とその枚数
各ゾーンのカードの枚数	自身のシールド，相手の手札などの確認不可能なゾーンのカードの枚数
デッキのカードの種類	自身が使用するデッキに含まれているカードの種類とその枚数
選択の内容	現在の進行中のアクション，選択するゾーンの種類のラベル
マナゾーンの文明	現在マナに含まれているマナの種類
行動中のカード	効果発動中や発動待機中のカードの種類．攻撃中や攻撃対象のカードの種類．
発動中の特殊効果	特定のカードによって発動している特殊な効果

表 2 行動の定義

選択対象	選択の内容
識別番号がn_1，状態sのクリーチャーカード	選択可能なクリーチャーカードの 1 つ．そのカードを対象にアクションや効果が発生する．ただし同じ識別番号のカードでも状態が異なるものは別々の選択として扱う．
識別番号n_2の呪文カード	選択可能な呪文カードの 1 つ．そのカードを対象にアクションや効果が発生する．
対戦相手	クリーチャーの攻撃時に相手を攻撃する．
選択対象なし	選択をスキップする．

5.4.2. 学習モデル

本研究では大まかに以下の流れで学習を行う．

① AI モデルの初期化．
② ゲーム環境の初期化．
③ ゲームの実行．DQN AI は行動を行うタイミングでゲーム環境から状態を受け取り，入力として扱えるように状態を処理する．学習中の AI はランダムと DQN を混ぜた形で行動を決定する(ε-greedy 法)．状態，行動，報酬は随時経験データとして記憶リストに保存される．勝利時と敗北時に報酬が得られる．
④ AI の学習．記憶リストに記憶データが一定数ある場合，記憶リストから経験データを一部取り出して学習を行う．②へ戻る．

6. 実験

6.1. デッキ選択モデルの性能評価

提案手法の評価実験の前にデッキ選択モデルの性能を評価する．そのため 10000 回の対戦で構築したデッキ選択モデルを用意し，それを使用するルールベース AI とランダムのデッキ選択モデルのルールベース AI を 50000 回対戦させた．表 3 はデッキ選択モデルの構築用に 10000 回対戦を行った際の各デッキの勝利数であり，表 4 は表 3 の結果を基にした各デッキの選択確率である．勝利数が多いデッキ 3 などは選ばれる確率が高くなり，勝利数が少ないデッキ 2 は選ばれる確率が低くなっている．表 5 はデッキ選択モデルを導入した AI の勝率である．それぞれのモデルは 0.515 から 0.522 程度の勝率であり，ランダムにデッキ選択を行う AI を僅かに上回っていた．

表 3 デッキ選択モデルの構築時の各デッキの勝利数

	1 回目	2 回目	3 回目
デッキ 1	1034	1033	1038
デッキ 2	878	852	829
デッキ 3	1047	1051	1088
デッキ 4	1082	975	995
デッキ 5	959	1007	1012

表 4 勝利数を基にしたデッキ選択の確率モデル

	1 回目	2 回目	3 回目
デッキ 1	0.2254	0.2716	0.2384
デッキ 2	0.0474	0.0431	0.0290
デッキ 3	0.2566	0.3261	0.3946
デッキ 4	0.3642	0.1506	0.1546
デッキ 5	0.1065	0.2085	0.1835

表 5 デッキ選択モデルを用いたルールベース AI の勝率

	1 回目	2 回目	3 回目
デッキ選択モデルありの AI の勝率	0.5155	0.5167	0.5223

6.2. 提案手法の評価実験

提案手法の導入でより良い AI の実装が可能かを実験で評価する．実験で用いる AI は一人回しを 20000 回，ルールベース AI を相手に対戦を 30000 回行って学習させた．

6.2.1. 報酬値の実験

提案手法の導入で効率的なカード使用を AI が学べるか評価するため，1 試行分の合計報酬値を平均した値を測定した．この実験は本研究の AI が学習用に一人回しを 20000 回行ったタイミングで行い，比較対象はランダム AI，ルールベース AI である．それぞれの AI は一人回しを 10000 回行い，結果は表 6 となった．測定値は概ねランダム AI が 0.78，ルールベース AI が 1.32 となり，DQN AI は 0.94 と 1.03 であった．一人回しを学習した DQN AI はルールベース AI には及ばないが，ランダム AI よりは高い値を出している結果になった．また図 3 は学習中に 1000 間隔で 1000 回一人回しを行い，その合計報酬の平均値の推移を表したものである（この時 AI は Q 値に従って行動する）．測定値は最初の 4000 回で急上昇をした後，0.8 から 1.1 の範囲の値を取り続けた．

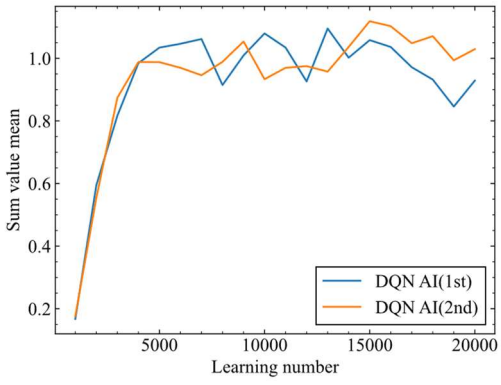

図 3 合計報酬の平均の推移

表 6 一人回し時の合計報酬の平均

	1 回目	2 回目
ランダム AI	0.7847	0.7834
ルールベース AI	1.3270	1.3281
DQN AI	0.9441	1.0390

6.2.2. 学習の評価

提案手法を導入した AI と導入していない AI の性能を比較するため，両方の AI でルールベース AI に対する勝率を測定した．この実験は本研究の AI がルールベース AI との対戦学習が終わったタイミングで行い，比較対象はルールベース AI を相手に 50000 回の対戦学習を行った AI である．実験はデッキ選択モデル（10000 回の対戦で構築）がある場合とない場合の両方で行い，勝率の測定のため各 AI はランダム AI，ルールベース AI と 10000 回の対戦を行った．結果は表 7，表 8 である．1 回目の一人回しありの AI は一人回しなしのいずれの結果も上回っていた．ただし 2 回目の一人回しありの AI はルールベース AI への勝率では一人回しなしの AI を上回っていたが，ランダム AI への勝率は下回った．また図 4，図 5 は学習中に 1000 間隔で 1000 回ルールベース AI と対戦を行い，その勝率の推移を表したものである（この時 AI は Q 値に従って行動する）．青線が通常の DQN AI，オレンジ線が一人回しありの DQN AI であるが，後者は前者と比べ早いペースで勝率が上昇し，ある程度学習が進むと勝率が追いついた．

表 7 ランダム AI に対する勝率

	1 回目	2 回目
一人回しなし（ランダムデッキ選択）	0.6738	0.6222
一人回しあり（ランダムデッキ選択）	0.6908	0.5972
一人回しなし（デッキ選択モデルあり）	0.7221	0.7365
一人回しあり（デッキ選択モデルあり）	0.7616	0.6822

表 8 ルールベース AI に対する勝率

	1 回目	2 回目
一人回しなし（ランダムデッキ選択）	0.1458	0.1074
一人回しあり（ランダムデッキ選択）	0.1627	0.1546
一人回しなし（デッキ選択モデルあり）	0.1814	0.1608
一人回しあり（デッキ選択モデルあり）	0.2148	0.1842

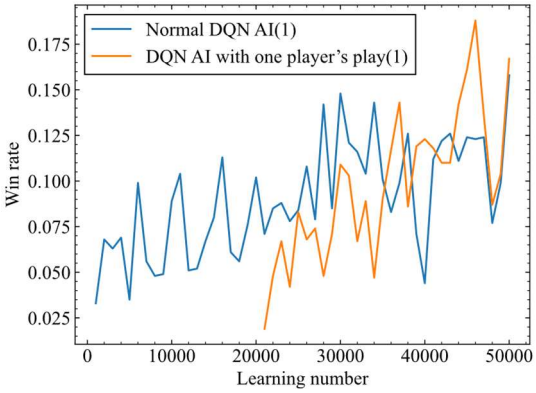

図 4 ルールベース AI への勝率 1

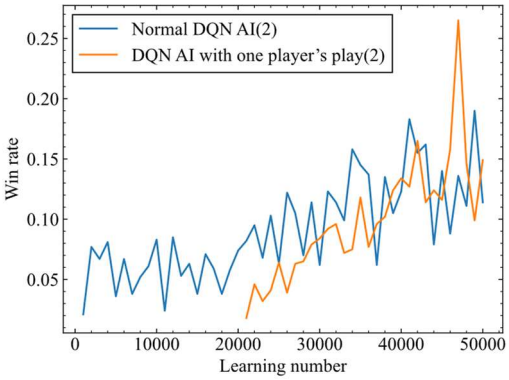

図 5 ルールベース AI への勝率 2

7. 議論

6.2.1 節の報酬値を用いた評価実験では，一人回しで学習した AI の合計報酬値の平均はランダム AI のものより高く，ルールベース AI のものより低い結果になった．この結果からルールベース AI と比較して適切なカードの使用は出来ていないものの，ランダム AI よりは測定値が高いため対戦相手に依存しない範囲のゲーム環境を多少学習出来ていることが分かる．また図 3 で示した一人回しの学習中に行った 1000 回毎のテストでは，最初の 4000 回の段階で数値の上昇が止まっている．この原因としては現在使用している学習のパラメーターがゲーム環境に対して合っていないことが挙げられる．一人回しと対戦学習で最適な学習のパラメーターが異なる可能性もあるため，各々で最適なパラメーターを見つける必要がある．

6.2.2 節の対戦による評価実験では，一人回しを導入して学習した DQN AI はノーマルの DQN AI と比べ，ランダム AI 相手の勝率は上回ったと言えないがルールベース AI 相手の勝率は全体的に上回っている．戦術的に行動を行うルールベース AI への勝率が提案手法の導入により上昇しているため，この結果から一人回しで効率的なカードの使用を学ぶことで AI の性能の向上が可能で，一人回しを導入することでより優れた AI を構築出来ると考える．ただし図 4，図 5 で示した対戦学習中に行った 1000 回毎のテストでは勝率の増減が激しく，学習が安定したものになっていなかった．この原因としては学習中の AI とその対戦相手のデッキが 5 種類あり各試合で使うデッキをランダムに決めていたため，1 試行毎に対戦の中身が大きく変わってしまっていたことが挙げられる．そのため本研究のようなカードゲームの学習を行う上では，ゲーム環境の変化に強い学習手法や多数のデッキを扱うための学習アルゴリズムを探すことが必要であると考える．また図 4，図 5 のグラフより学習の終了時付近でも勝率が大きく変動しており学習が収束している様子がなかった．そのため一人回しの効果をより明確にするには学習回数を増加させる必要もある．また本研究で実装した DQN AI は全体的にルールベース AI に対する勝率が低くなっている．その原因としては，学習モデルがデュエルマスターズの複雑なゲーム環境全体を扱えず学習が不完全になったことも挙げられる．一人回しの対戦モデルへ

の効果をより明確にするにはこうした問題も解決する必要があると考える．

デッキ選択モデルに関しては，6.1 節の実験でデッキ選択モデル導入した AI はそうでない AI を僅かに上回った．6.2.2 節の実験ではデッキ選択モデルの導入により 3～11%近く勝率を上げており，その効果が見られた．デッキ選択モデルの導入はあまり勝てないデッキの使用頻度を下げ，よく勝てるデッキの使用頻度を上げられ，それにより全体的な勝率を上げられたと考える．また 6.2.2 節の結果については，学習過程上各デッキに学習度合いのばらつきがあり，デッキ選択モデルによってよく学習出来たものを優先に使用したため勝率の上昇が見込めた可能性もあると考えた．

8. おわりに

本研究ではカードゲーム「デュエルマスターズプレイス」を題材に深層強化学習を用いた自動対戦 AI の実装を行った．その複雑なゲーム環境を学習するために「一人回し」を導入した．一人回しは対戦相手がいない状況で学習を行い，カードの使用に応じて報酬がもらえるため，対戦相手に依存しない範囲のカードプレイなどを学べる．本研究で行った実験から，一人回しで学習することである程度効率的なカードの使用を学習出来ること，また一人回しを導入している AI の方が導入していない AI より学習出来ることが分かった．

今後の課題として，基盤となる AI がルールベース AI にあまり勝てていないため，DQN 以外の学習手法を用いて実験を行うこと，学習手法以外にも各種パラメーターやネットワークモデルを改良して実験を行うことが挙げられる．加えて，学習速度などの観点から十分な学習環境であるとは言えないため，GPU を用いた並列での学習を導入することが挙げられる．

文　献

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra and M. Riedmiller, "Playing Atari with Deep Reinforcement Learning," arXiv preprint arXiv:1312.5602, 2013.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, et al., "Human-level control through deep reinforcement," *Nature,* vol. 518, no. 7540, pp. 529-533, 2015.

[3] W. Xi, Y. Zhang, C. Xiao, X. Huang, S. Deng, H. Liang and J. Chen, "Mastering strategy card game (legends of code and magic) via end-to-end policy and optimistic smooth fictitious play," arXiv preprint arXiv:2303.04096, 2023.

[4] C. Xiao, Y. Zhang, X. Huang, Q. Huang, J. Chen and P. Sun, "Mastering strategy card game (hearthstone) with improved techniques," *2023 IEEE Conference on Games (CoG),* pp. 1-8, 2023.