カードゲーム型対戦環境への深層強化学習の適用

1　はじめに

近年, 人工知能に関する研究分野は目覚ましい発展を遂げており様々な分野に応用されている. その中でも人間の学習プロセスに近いとされる強化学習と深層学習を融合した深層強化学習が注目されている.

深層強化学習の応用先の 1 つとしてゲームがある. 特に将棋や囲碁といった, プレイヤーが意思決定をする段階でそれ以前の意思決定の過程がすべて把握可能な完全情報ゲームへの応用において成果が顕著である. 最近では麻雀やポーカーのような, プレイヤーに与えられる情報が部分的である不完全情報ゲームへの応用も注目されている.

そこで, 本研究では不完全情報ゲームであるトレーディングカードゲームを参考にカードゲーム型対戦環境を構築し, 構築環境へ深層強化学習手法を適用し学習済みエージェントの行動を観測した.

2　要素技術

2.1　OpenAI Gym

OpenAI Gym [1] は非営利企業 OpenAI が提供する強化学習のシミュレーション用ライブラリであり, 強化学習の環境として多くのゲームが登録されている. 提供されているインターフェースに沿って, エージェントの行動空間や状態空間, 報酬などを定義, 実装することで自作の環境を登録し利用することができる. 様々な強化学習用ライブラリに対応しているため比較的容易に強化学習を試すことができる.

2.2　Deep Q Network

価値ベースの強化学習手法の 1 つ である Q 学習では実装の際に状態と行動をインデックスとする Q 値のテーブルを作成する. しかし状態空間や行動空間が高次元である, あるいは状態や行動が離散値ではなく連続値で表現される場合には Q テーブルのメモリ量は爆発してしまう. この問題を解決した技術が Deep Q Network (DQN) [2] である. DQN ではニューラルネットワークを用いて, ある状態における行動ごとの Q 値を推定する. エージェントが経験した過去の体験を Replay Memory に一定期間保存しておき, 過去の経験をランダムにサンプリングして学習する Experience Replay や行動を決定するネットワークと Q 値を学習するネットワークを分けることで Q 値の過大評価を防ぐ Fixed Target Network といった工夫により安定した学習を可能としている.

3　カードゲーム型対戦環境の構築

3.1　トレーディングカードゲーム

本研究で構築したカードゲーム型対戦環境は, Magic : The Gathering [1] といったトレーディングカードゲーム (Trading Card Game : TCG) を参考にした. TCG は 2 人のプレイヤーからなるゲームである. プレイヤーは先攻と後攻に分かれ, ターン制で進んでいく. 大きな特徴として将棋やチェスのように同じユニットを用いるのではなく, 事前に各プレイヤーの選択による異なるユニットからなるデッキを構築する点が挙げられる. ゲームタイトルごとに異なるが, 多くの場合相手プレイヤーのカードの 1 部分はプレイヤーから観測できない不完全情報ゲームである.

3.2　構築環境

実装したカードゲームのルールと用語を説明する. ゲームは 2 人のプレイヤーからなり, プレイヤーは複数のカードからなるデッキを持つ. また, HP , マナといった 2 つの数値を持つ. プレイヤーは手札, 盤面と呼ばれるカードを保有する領域を持ち, ドローと呼ばれる操作でカードをデッキから手札に加える. また, プレイと呼ばれる操作でカードを手札から盤面に出す. プレイヤーがカードをプレイする際, 後述するカードのコスト分マナを消費する. また, デッキからカードが無くなった状態をデッキ切れと呼ぶ. 今回の環境ではプレイヤーの HP は 20 , マナの上限値は初期値が 1 で最大値を 5 とした.

カードはそれぞれ攻撃力と HP とコストの 3 つの数値を持つ. また, カードによっては特殊効果を持つものもある. 盤面にあるカードは対戦相手の盤面にあるカード, あるいは相手プレイヤーに攻撃することができる. ただし, 攻撃が可能となるのはカードがプレイされたターンの次のターンからになる. カードが攻

[1] https://magic.wizards.com

撃する際には, 相手盤面に存在する攻撃対象のカードの HP, あるいは相手プレイヤーの HP へカードの持つ攻撃力分ダメージを与える. またカードへと攻撃する際には攻撃対象のカードが持つ攻撃力分, 攻撃するカードもダメージを受ける. カードの HP が 0 になった, あるいは後述する手札と盤面の枚数制限を超えて盤面にプレイされた時はカードは破壊される. 破壊されたカードはゲームから取り除かれる.

3.3 ゲームフロー

実装したゲームの流れを説明する.

1. ゲーム開始時に各プレイヤーは自身のデッキをシャッフル.
2. デッキから初期手札としてカードを 5 枚ドロー.
3. 先攻プレイヤーは 1 ターン目のドローステップをスキップし行動.
4. 後攻プレイヤーはカードを 1 枚ドローして行動.
5. 2 ターン目以降は先攻プレイヤーもカードを 1 枚ドローしてから行動.
6. 4 , 5 の繰り返し. なお, ターンプレイヤーは行動前にマナを上限値まで回復. このときマナの上限値が 5 でなければ上限値を 1 増やしてから回復.
7. プレイヤーがデッキ切れになっている状態でカードをドローしようとした, あるいはプレイヤー自身の HP が 0 となった場合はそのプレイヤーが敗北となりゲーム終了.

本構築環境では, 一般的な TCG と同様に先攻プレイヤーがカードの行動が早いため有利となる. そのため, 先攻の 1 ターン目のドローステップをスキップしている.

4 実験

構築環境へ深層強化学習が適用できるか検証した. 深層強化学習手法として DQN を用いて構築環境において先攻のプレイヤーとして学習し, 学習済みのエージェントで 10000 回ゲームを実行して勝率を計算した. また, 学習が進んでいるかどうか判断するため学習時の獲得報酬の推移を記録した. さらに 50000 回の対戦においてエージェントが選択した行動の総数, 各カードごとのプレイされた総数を記録し学習済みのエージェントがどのような戦略を構築したか, それが人間から見て合理的な戦略かどうか考察した.

4.1 対戦相手の行動ルーチン

学習, 勝率計算の際には学習するプレイヤーの対戦相手を用意する必要がある. Algorithm 1 に今回の実験における対戦相手の手動で作成した行動ルーチンを示す.

Algorithm 1 対戦相手の行動ルーチン

```
盤面にカードを 1 枚プレイ
for 盤面のカード (プレイ順が古い方から) do
    if 敵の盤面に 1 回の攻撃で破壊できるカードがある then
        その攻撃対象を選んで攻撃
    else
        if 敵盤面の総攻撃力が自身の HP 以上 then
            敵盤面の最も攻撃力高いカードを攻撃
        else
            敵プレイヤーを攻撃
        end if
    end if
end for
ターンを終了
```

表 1: デッキに採用したカードの内容

ID	攻撃力	HP	コスト	特殊効果
0	1	1	0	無し
1	2	1	1	無し
2	3	2	2	無し
3	4	3	3	無し
4	5	4	4	無し
5	2	2	2	召喚
6	2	3	3	召喚
7	1	1	1	取得
8	1	3	2	取得
9	2	1	2	速攻
10	3	1	3	速攻
11	1	2	2	攻撃
12	2	3	3	攻撃
13	1	1	1	治癒
14	2	1	3	治癒

4.2 デッキ

学習側, 対戦相手ともに同じデッキを持つ. 表 1 に, デッキに採用したカードの内容を示す. デッキは 表 1 に示すカードを各 2 枚ずつ, 計 30 枚のカードから構成される. 特殊効果は便宜上 2 字の単語で表しており, 具体的な効果は以下の通りである.

召喚： 盤面に出したら (攻撃力 , HP) = (1 , 1) のユニットを追加で盤面に出す

治癒： 盤面に出したら自プレイヤーの HP を 2 回復する

攻撃： 盤面に出したら敵プレイヤーの HP を 2 減らす

取得： 盤面に出したら自プレイヤーは 1 枚カードをドローする

速攻： 盤面に出たターンに攻撃できる

表 2: 定義した状態空間

状態説明	次元数	最小値	最大値
各プレイヤーの HP	2	0	20
各プレイヤーの マナ	2	0	5
手札 1 ～ 9 の HP , 攻撃力, コスト, 特殊効果	36	0	5
自盤面 1 ～ 5 の HP と攻撃力	10	0	5
敵盤面 1 ～ 5 の HP と攻撃力	10	0	5
自盤面 1 ～ 5 が 攻撃可能かどうか	5	0	1
お互いのデッキの 残り枚数	2	0	30

表 3: 定義した行動空間

行動説明	次元数
手札 1 ～ 9 を自盤面に出す	9
自盤面 1 が敵盤面 1 ～ 5 に攻撃 or 敵プレイヤーに攻撃	6
自盤面 2 が敵盤面 1 ～ 5 に攻撃 or 敵プレイヤーに攻撃	6
自盤面 3 が敵盤面 1 ～ 5 に攻撃 or 敵プレイヤーに攻撃	6
自盤面 4 が敵盤面 1 ～ 5 に攻撃 or 敵プレイヤーに攻撃	6
自盤面 5 が敵盤面 1 ～ 5 に攻撃 or 敵プレイヤーに攻撃	6
ターンエンド	1

4.3 状態空間と行動空間, 報酬の定義

強化学習では, エージェントの取りうる行動と観測できる状態の空間, 報酬を定義する必要がある. TCG ではドローやプレイ, カードの攻撃による破壊といった行動で盤面や手札の枚数が変化する場合があり, 各ステップ時点でプレイヤー取りうる行動の次元数が変わるため学習が困難である.

そのため本研究では予め手札と盤面の枚数の上限をそれぞれ 9 枚, 5 枚と定め, 手札と盤面に存在するカードに自盤面 1 というように番号をつけ, カードが存在しない場合は状態を 0 とすることで状態空間と行動空間を定義した. 表 2, 3 に状態空間, 行動空間の定義を示す. なお, ドローやプレイといった操作でカードを追加し枚数の上限を超える場合には追加しようとしたカードを破壊する.

報酬は以下のように設定した.

- 1 ステップ終了後

$$\mathrm{r} = 0.0$$

- 1 エピソード終了後

$$\mathrm{r} = \begin{cases} 1.0 & \text{(学習プレイヤーの勝利)} \\ -1.0 & \text{(敵プレイヤーの勝利)} \end{cases}$$

4.4 DQN のパラメータ

表 4 に実験で用いた DQN のパラメータを示す.

表 4: DQN のパラメータ

パラメータ名	値
割引率 γ	0.99
全結合層の活性化関数	ReLU
全結合層の次元	64
最適化アルゴリズム	Adam
方策	ε -greedy
Target Network 更新重み	0.5
Exprience Memory 開始ステップ数	1.0×10^5
学習ステップ数	1.0×10^6

表 5: 実験結果 (後攻プレイヤーは同じ)

先攻プレイヤーの戦略	勝率
DQN で学習済み	**0.9739**
対戦相手と同じ行動ルーチンで行動	0.5089
表 3 の行動空間に沿ってランダムに行動	0.3155

また, ε -greedy において ϵ は (1) 式に従って, $\epsilon_{\max}$ から $\epsilon_{\min}$ へと指数関数的に減少する.

$$\epsilon = \max(\epsilon_{\min}\ ,\ \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min}) \exp(-\frac{\mathrm{n_{step}}}{\epsilon_{\mathrm{decay}}})) \quad (1)$$

ここで $\mathrm{n_{step}}$ は学習時の累積ステップを表す. 本実験では $\epsilon_{\max} = 1.0, \epsilon_{\min} = 0.1, \epsilon_{\mathrm{decay}} = 50.0$ とした.

5 結果と考察

表 5 に実験結果を示す. なお, ベースラインとして Algorithm 1 に示す行動ルーチンで動くプレイヤー, 表 3 の行動空間に沿ってランダムに行動するプレイヤーを先攻に配置して 10000 回対戦を実行し勝率を計算した結果を示している.

DQN で学習したエージェントは, 9 割 7 分とベースラインと比べて遥かに高い勝率を残している. また, 図 1 に学習時の平均獲得報酬の推移を示す. 図 1 において縦軸は reward, 横軸はエピソード数であり, 図中の緑点は 学習時の 500 エピソードにおける平均獲得報酬を表し, 薄緑の領域は標準偏差を表す. およそ 2000 エピソード学習した段階で獲得報酬が大きく上昇し, 5000 エピソードを超えたあたりで学習が安定している. また, 表 6 に 50000 回の対戦においてランダムに行動するプレイヤーが選択した総数が多い上位 5 つの行動と学習済みエージェントが選択した総数が多い上位 5 つの行動を示し, 表 7 に 50000 回の対戦において各カードがプレイされた回数を示す. (1) 式から, 学習序盤はエージェントはランダムに探索をしているためランダムに行動するプレイヤーは学習序盤のエージェントが取る行動とみなせる. そのため学習済エージェントの対戦データの比較対象としてランダムに行動するプレイヤーの対戦データを用いる.

表 6: 選択された総数が多い行動上位 5 つ

ランダム		DQN	
行動説明	総数	行動説明	総数
ターンエンド	596214	手札 1 を盤面にプレイ	386541
手札 1 を盤面にプレイ	344363	ターンエンド	309775
手札 2 を盤面にプレイ	225676	盤面 1 で相手プレイヤーに攻撃	262587
盤面 1 で相手プレイヤーに攻撃	149547	盤面 2 で相手プレイヤーに攻撃	139880
盤面 1 で敵盤面 1 を攻撃	122671	手札 2 を盤面にプレイ	76888

表 6 において, 学習済エージェントではランダムに行動するプレイヤーと比べると各行動の総数の数値が全体的に減少しており, また相手プレイヤーに直接攻撃する行動が多く選択されていることが分かる. このことから, DQN では学習により相手プレイヤーに直接攻撃して早くゲームを終了させる戦略が最適と学習されたといえる. また, ランダムに行動するプレイヤーが最も多く選択していたターンエンドは, DQN では 2 番目まで下がっている. これは, まだカードがプレイできる状態でターンエンドするといった無駄なターンエンドをする傾向が減っていることを意味する. これは人間から見ても妥当な行動といえる.

表 7 において, DQN で学習済のエージェントは 0, 1, 2 コストのカードに関してはコストが小さいカードほど多くプレイされている. これは先述したような無駄なターンエンドは損であると学習し, プレイヤーのマナを余らせること無くカードを盤面にプレイできていると考えられる. また, 4 コストの ID 4 のカードが 3 コストの ID 14 に比べて多くプレイされている. ID 4 のカードは HP が唯一 4 となるカードであり対戦相手は, Algorithm 1 に従うため, 対戦相手の HP が自盤面の総攻撃力以下である場合と, 対戦相手が ID 3, 4 のカードを盤面に出す場合以外では ID 4 のカードは破壊されることがない. また攻撃力も 5 と全カード中最大であるため本実験条件においてはカードパワーが高いカードである. 一方で, ID 14 のカードは 3 コストのカードの中で攻撃力と HP の和が最も低く, 特殊効果も治癒であるため相手プレイヤーの HP をなるべく早く減らすといった戦略とは親和性が低い. そのため ID 14 のカードを盤面にプレイするよりも, コストが 1 高いが強力な ID 4 のカードを盤面に出すことが得であると学習している. すなわち構築した戦略において各カードの強弱も学習しており, その結果も人間から見て合理的な判断となっている.

6　まとめと今後の課題

今回の実験では TCG を参考としたカードゲーム型対戦環境を構築した. 構築環境へ DQN を適用し, 対戦相手に対して高い勝率を記録し, 人間から見ても合理的な戦略をもつエージェントを作成した.

今回の環境では相手の盤面のカードに攻撃するよりも相手プレイヤーを直接攻撃する行動が最適と学習により判断された. そのため手札から最もダメージが期待できるカードを出すだけになってしまい, 盤面における戦略性が無くなってしまった. デッキの構成や対戦相手の戦略, 新たな特殊効果の追加などでより複雑で戦略性の高いカードゲーム型対戦環境を構築することが今後の課題として挙げられる.

表 7: 各カードが盤面にプレイされた総数 (降順)

ランダム		DQN	
ID	総数	ID	総数
0	60624	0	42462
1	59675	1	40223
9	59310	13	40129
11	59198	7	39376
13	58762	11	38578
2	58075	9	38402
5	57498	5	38112
10	57288	2	38019
7	57204	8	37703
12	56148	3	34682
8	55872	12	34296
14	55167	10	33971
3	55079	6	33840
6	53809	4	33778
4	53745	14	32522

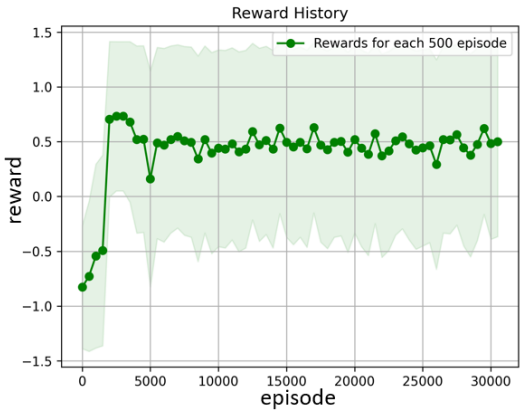

図 1: DQN における平均獲得報酬の推移

参考文献

[1] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv e-prints*, p. arXiv:1606.01540, June 2016.

[2] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. 2013. cite

arxiv:1312.5602Comment: NIPS Deep Learning Workshop 2013.