2022 年 12 月 13 日 B3 西村昭賢

進捗報告

1 今週やったこと

- 研究発表会の準備 (スライド)
- ゲームバランス調整のアプローチ検討
- 自作環境の改良 & ルール確定
- エージェント作成実験

2 ゲームバランス調整について

構築環境への強化学習手法の適用を研究の目的の 1 つとしたため, 強化学習で作成したエージェントを用いてバランス調整を行いたいと考えた.

インターンでゲーム開発現場においてパラメータチューニングが属人化しやすいという話を聞いたので, バランス調整ではカードの HP, 攻撃力を調整することにする. 現在は図 1 のようなアプローチを考えている.

- 良いと考えている点
  - 強化学習でルールベース AI に勝つような戦略を持つ AI を作成することでより実践的なシミュレーションが行える.
  - 調整したいデッキを用意することで自動で行える
- 良くないと考えている点
  - 深層強化学習を用いると学習に時間がかかる
  - ルールベースに勝つように学習するためルールベースで作成した AI に大きく依存する

各カードの戦績の計算に用いる指標としては,

- 盤面に出されてから生きているターン数
- 倒した敵カードの数
- 敵プレイヤーに与えたダメージ

を考えている. また, HearthStone のバランス調整を試した研究 [1] では Win Rate when Played (WRP) , Win Rate when Drawed (WPD) という指標を用いられている.

まだ考えがまとまっていないが, 強化学習の学習済みエージェントでシミュレーションすることでより実践的な対戦データを得られるためバランス調整に生かせそうと考えている.

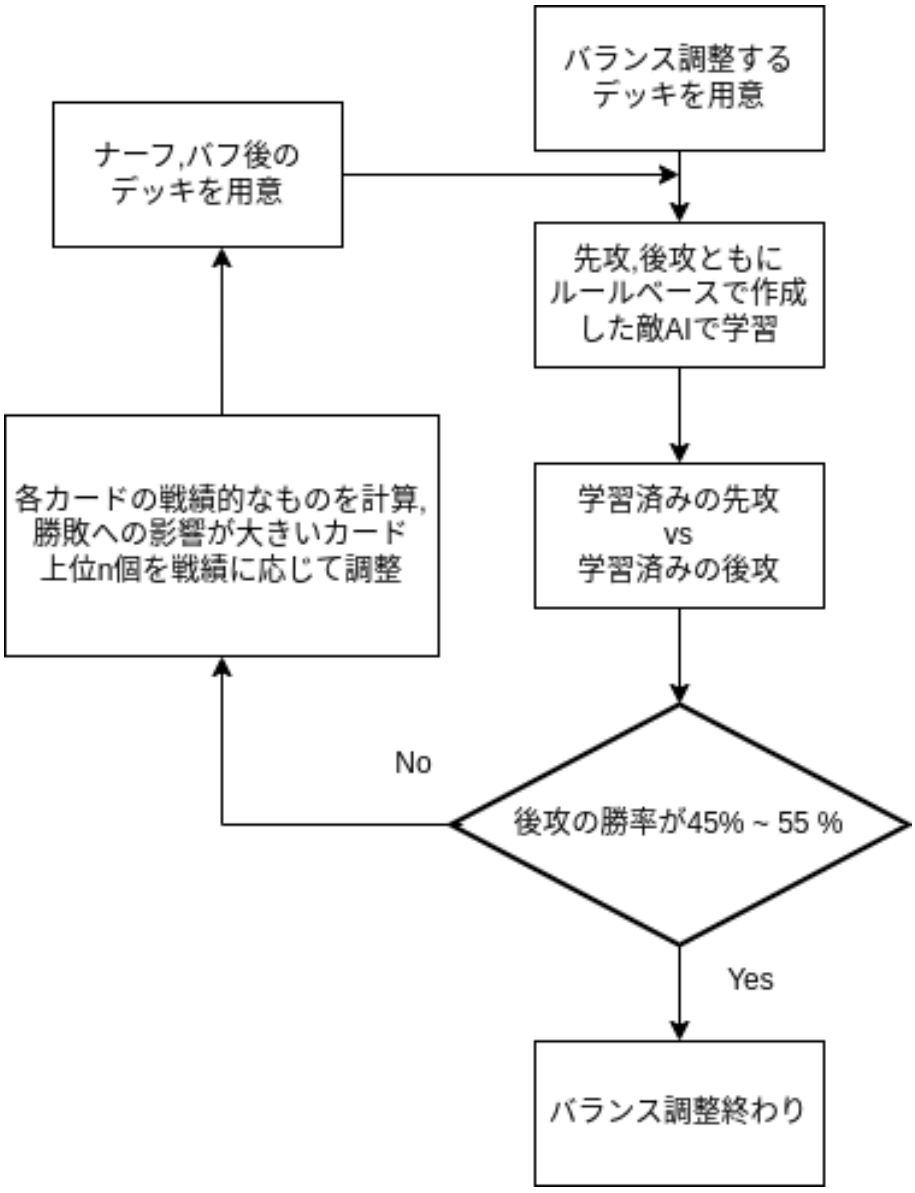

図 1: バランス調整アイデア

3　自作環境の改良 & ルール確定

バランス調整に取り掛かるにあたって, 先週までのルールはコストがなくカードのプレイに関して戦略性がないルールになっていた. より一般的な TCG に寄せたルールのゲームを作成し, 構築環境として確定しバランス調整に取り組みたいと考えた. 追加した点をいかに述べる.

- プレイヤー
  - HP
    最大 20, 0 となればゲーム敗北
  - マナコスト
    ゲーム開始時 1 , 最大 5 , ターンごとに 1 増加
  - ライブラリ
    ライブラリは 30 枚のカードを持つ

- カード
  - コスト
    盤面にプレイする際にカードのコスト分プレイヤーのマナコスト減少
  - 特殊効果
    - 盤面に出したら (攻撃力 , HP) = (1 , 1) のユニット追加で出す. (召喚)
    - 盤面に出したら自プレイヤーの HP を 2 回復　(治癒)
    - 盤面に出したら敵プレイヤーの HP を 2 削る　(攻撃)
    - 盤面に出したら自プレイヤーは 1 枚カードをドロー　(循環)
    - 盤面に出たターンに攻撃できる (速攻)
- 終了条件
  どちらかのプレイヤーの体力が 0 以下となった , または デッキ切れの状態でドローしようとした時

4　実験

強化学習が適用できるか実験を行った.

4.1　行動空間と状態空間の定義

環境の改変により定義し直した状態空間と行動空間は表 1 , 2 に示す.

表 1: 定義した状態空間 (太字は新しく追加したパラメータ)

状態説明	次元数	最小値	最大値
自, 敵プレイヤーの HP	**2**	**0**	**20**
自, 敵プレイヤーのコスト	**2**	**0**	**5**
手札 1 ∼9 の HP と攻撃力	18	0	20
手札 1 ∼ 9 のコスト	**9**	**0**	**5**
手札 1 ∼ 9 の特殊効果	**9**	**0**	**5**
自盤面 1 ∼ 5 の HP と攻撃力	10	0	20
敵盤面 1 ∼ 5 の HP と攻撃力	10	0	20
自盤面 1 ∼ 5 がターン中行動可能かどうか	5	0	1
お互いのライブラリの残り枚数	2	0	15

表 2: 定義した行動空間 (太字は今回新規に追加したパラメータ)

行動説明	次元数
手札 1∼9 を自盤面に出す	9
手札 1∼9 を自盤面に出さない	9
自盤面 1 が敵盤面 1∼5 に攻撃 or 何もしない or **敵プレイヤーに攻撃**	**7**
自盤面 2 が敵盤面 1∼5 に攻撃 or 何もしない or **敵プレイヤーに攻撃**	**7**
自盤面 3 が敵盤面 1∼5 に攻撃 or 何もしない or **敵プレイヤーに攻撃**	**7**
自盤面 4 が敵盤面 1∼5 に攻撃 or 何もしない or **敵プレイヤーに攻撃**	**7**
自盤面 5 が敵盤面 1∼5 に攻撃 or 何もしない or **敵プレイヤーに攻撃**	**7**

4.2 ライブラリ

プレイヤーのライブラリは先攻と後攻同じものとした. デッキの詳細を表 3 に示す.

表 3: ライブラリ () 内の数字は (攻撃力, HP, コスト) を表す

攻撃力	HP	コスト	特殊効果	枚数
1	1	0	無し	2
2	1	1	無し	2
3	2	2	無し	2
4	3	3	無し	2
5	4	4	無し	2
2	2	2	召喚	2
2	3	3	召喚	2
1	1	1	循環	2
1	3	2	循環	2
2	1	2	速攻	2
3	1	3	速攻	2
1	2	2	攻撃	2
2	3	3	攻撃	2
1	1	1	治癒	2
2	1	3	治癒	2

4.3 DQN のパラメータ

DQN のパラメータを表 4 に示す. ε-greedy について学習が進むに連れて学習率が減少していくように変更した.

表 4: DQN のパラメータ

方策	ε-greedy
ε の初期値	1.0
ε の最小値	0.1
ε の推移	(学習ステップ数 / 3) ステップまで線形的に減少
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
Target Network 更新重み	0.5
Exprience Memory への書き込み開始 step	10000
Experience Replay のメモリ量	50000

4.4 敵プレイヤーの行動

ルールベースである程度強い AI 相手と学習させたかったので以下のような行動ルーチンとした.

Algorithm 1 敵の行動

```
for 手札のカード do
    if 盤面にプレイできる then
        カードをプレイ
    else
        pass
    end if
end for
for 自盤面のカード do
    if 敵の盤面に 1 会の攻撃で倒せるカードがある then
        そのカードを選んで攻撃
    else
        敵プレイヤーを攻撃
    end if
end for
```

要するに手札から出せる分出して, 敵盤面に倒せるカードがあればカードを攻撃, なければ敵プレイヤーを攻撃するという攻撃的な行動になっている. この敵を対象に DQN で 先攻プレイヤーを 1000000 ステップ学習を行い, 学習済みモデルで 10000 回対戦を行い勝率を計算した.

4.5 報酬

報酬は (1) 式 のように設定した.

$$reward = 0.0, \quad 1\text{エピソード終了後}\, reward = \begin{cases} 1.0 & \text{(学習プレイヤーの勝利)} \\ -1.0 & \text{(敵プレイヤーの勝利)} \end{cases} \tag{1}$$

4.6 結果と考察

実験結果を表 5 に示す. また, 図 に学習中の DQN の 200 エピソード中の獲得報酬平均の推移を示す

表 5: 実験の結果

手法	勝率
敵プレイヤーと同じ行動	**0.5033**
DQN	0.1461

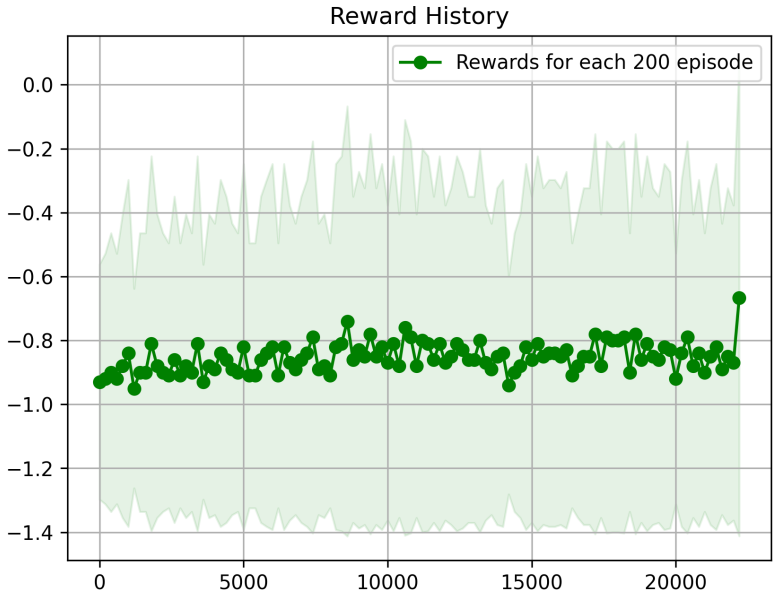

図 2: DQN 学習中における獲得報酬平均の推移

結果としては DQN で 1 割程度の勝率しか得ることができなかった. 学習が上手く進まなかった原因として以下の要因が考えられる.

- 学習ステップ数不足
  今回の実験では対戦相手がルールベースの AI であり reward が 1 となるエピソードが少ない. そのため学習が上手く進まなかったと考えられる. 今回はゼミに間に合わせるために 1000000 ステップ, 約 22000 エピソード学習を行ったが, ステップ数を増やして実験を試してみる.
- reward の与え方
  一般的な TCG の勝利条件は相手のプレイヤーの HP を 0 にする, または相手がデッキ切れ起こすことであり, 今回作成した環境でも同じ勝利条件を採用した. よく考えてみると相手のプレイヤーの HP を 0 にするようにはある程度攻撃的に行動しなければならず, 相手のデッキ切れを誘うには防戦的に動く必要がある.
  今回の実験では reward を　勝利すれば 1.0 としていたため学習側プレイヤーの戦略が学習により上手く構築できなかったのではないかと考えた. 例えば相手の HP を 0 にしたら reward = 1.0, 相手のデッキ切れで勝ったら reward = 0.1 など重みをつけたら学習の進み具合が変わるかもしれない.

5　わからんこと

今回, DQN で ε-greedy の ε を最初は 1.0 として 0.1 に収束させる方法を採用したのですが, ε の減少の方法が調べてもあまり見つかりませんでした. 今回の実験では (学習ステップ数 / 3) ステップまで線形的に減少としましたが , なにか良い方法があればご教授いただければ幸いです.

6　今後やること

- 研究発表会の準備 (資料作成)

- ゲームバランス調整のアプローチ検討
- 改良した構築環境への強化学習適用
  先週までのゲームルールから戦略性, ゲーム性を増したいだけなのでプレイヤーの HP 有りで上手く行かなかったら HP 無しに切り替える予定. なるべく HP 有りでやってみたいので今週はいろいろ検討してみる.

参考文献

[1] Fernando de Mesentier Silva, Rodrigo Canaan, Scott Lee, Matthew C. Fontaine, Julian Togelius, and Amy K. Hoover. Evolving the Hearthstone Meta. *arXiv e-prints*, p. arXiv:1907.01623, July 2019.