進捗報告

1　やったこと

- 自己対戦 (self-play) の実装
- 今後の方針

2　自己対戦 (self-play) の実装

JSAI で発表されたテーマの中で、ぷよぷよ [1] と逆転オセロニア [2] では自己対戦を用いていたため、実装してみた. エージェントに持たせるデッキは前々から用いている恣意的に強いカードと弱いカードを入れたデッキを用いている. 表 1 に DQN のパラメータを示す. また、図 1 に学習時の報酬の推移を示す. ここで、reward は先攻側の勝敗で決定している。さらに表 2 に先攻にアグロとコントロールでランダムに戦略が変化するエージェントをおいて後攻に学習済みエージェントを配置したときの学習済みエージェントを示す。

うまくいかなかった原因として、報酬の設計が考えられる。何も考えずに専攻の勝敗で報酬を設計してしまったため、今エージェントがプレイしているサイド (先攻後攻どちらか) から見た立場で勝敗から報酬を設計してやってみたい。また、逆転オセロニアの場合では自己対戦を用いているが、その前に以前のバージョンのエージェントの戦略とトッププレイヤーのログから教師あり学習を行っている。また、自己対戦ではないがスタークラフト 2 を対象としている AlphaStar [3] においても強化学習の前にトッププレイヤーのログからオフライン強化学習をしている。報酬が勝敗のみで与えられる環境においては、自己対戦だと学習までの立ち上がりが遅いことが考えられるためこういった手法が必要かもしれない。

表 1: 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 2: 後攻の戦略を変化させた場合の勝率比較

後攻の戦略	勝率
DQN 学習済エージェント	**0.7182**
selfplay 学習済みエージェント	0.5141
行動空間に沿ってランダム	0.2336

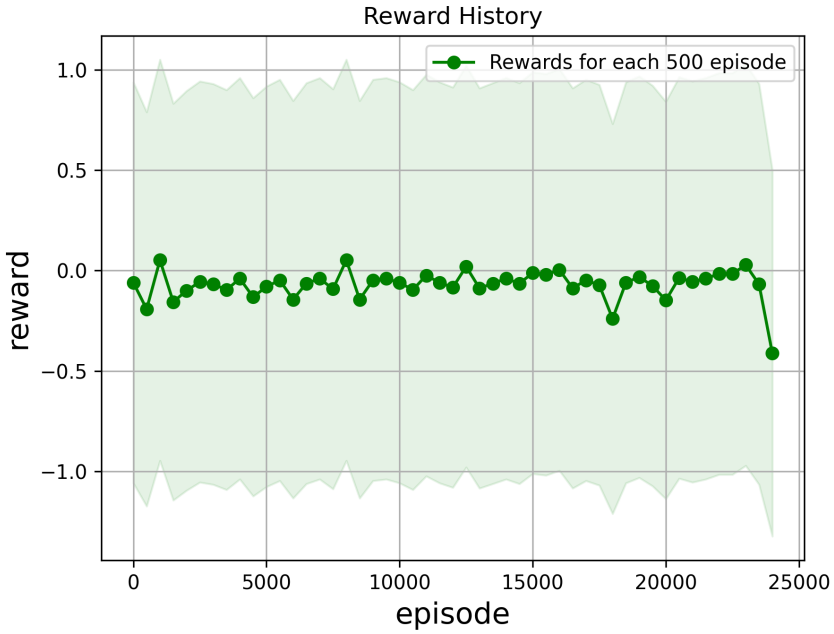

図 1: self-play における 500 エピソードごとの平均獲得報酬の推移

3 今後の方針

- テーマに関して

  岩倉さんから新しい環境を頂けることを期待しています。マルチモーダルなデータから強化学習試してみたい。頂けなさそうな場合はまだ考えていません。

- 新しい深層強化学習アルゴリズムの実装

- カードプールからデッキの構築

  事前にデッキを決めていることに関しての指摘が多く、以前から取り組んでみたかったため検討してみたが、２つの懸念点がある。

  – 事前に戦略決めないとデッキ構築の方針が立たない、方針を事前に与えたとしても評価どうするか

  – カードプールのパラメータの設定どうするか (自分が設定するのもなんかな〜って思ってます)

参考文献

[1] 福地 昂大、三宅 陽一郎. 『ぷよぷよ』における深層強化学習による自己対戦の適応. https://confit.atlas.jp/guide/event/jsai2023/subject/2M5-GS-10-01/date?cryptoId=.

[2] 佑甲野, 一樹田中, 健岡田, エルネスト 純奥村. “逆転オセロニア”における深層強化学習応用. デジタルプラクティス, Vol. 10, No. 2, pp. 351–367, jan 2019.

[3] Kai Arulkumaran, Antoine Cully, and Julian Togelius. Alphastar: An evolutionary computation perspective, 2019. cite arxiv:1902.01724.