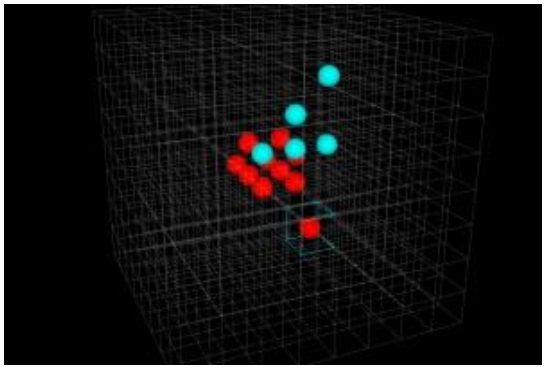
DQNを用いた3次元リバーシAIの特徴の考察



メンバー

185732B 船附海斗

185734H 高良　隼

185743G 大城祐介

185751H 長濱北斗

1.目的・目標

目的:

　ゲームを通して強化学習を学ぶ。

目標:

　DQN法のハイパーパラメータの変化によって勝率がどう変化するか、DQN法の特徴を考察する。また3Dリバーシにて高性能なアルゴリズムを設計する。

2.アプローチ全体像（1）

　3Dリバーシを用意し、DQN法を適用させ、ランダムに手を打つプレイヤーと戦わせる。その際、ハイパーパラメーターを変化させていくことによって勝率のデータを収集する。それをもとにDQN法の特徴を考察する。

2.アプローチ全体像（2）

3Dリバーシと DQN法を選択した理由について

- 3Dリバーシにした理由
  2Dリバーシより複雑な学習環境であるので戦略の幅が広く、DQNによる自律的な特徴量の抽出とモンテカルロ木探索によるヒューリスティックな探索が良い勝負をすると考えた。

- DQN法を選んだ理由
  先行研究が多数あり、情報も豊富なため参考にしやすいため。

3.予定していた実験計画

I. 2D、3Dのリバーシの用意、強化学習の勉強

II. 使用するアルゴリズムの選択

III. アルゴリズムの実装

IV. ハイパーパラメータの調整

V. 各アルゴリズムの比較

VI. アルゴリズムの特徴を考察

4.データセットの構築方法

- 戦績データ
  - s: 盤面データ
  - a: AIが石を配置する場所
  - s2: AIがsに対してaの位置に石を配置した後の盤面の状態
  - r: sからs2に遷移するときに得た報酬
    - (勝利=1,敗北=-1,引き分け=-0.5)
  - t:ゲームが終了したか(if 終了 t = 1; else t = 0;)
- ゲームを進めながら過去の戦績データを最大1万個分までリストに保存し、ランダムに複数個取り出して学習を行う

5.機械学習の進め方（1）

DQN法について

- 行動価値関数(Q値)をニューラルネットを使った近似関数で求め、ある状態の時に行動ごとのQ値を推定し、取るべき最善の行動を求める手法です。
- 活性化関数にLeaky_ReLUを利用したMLP(多層パーセプトロン)でDQNを作成しました。
- 現在の盤面状態ととりうる手を入力に、行動価値関数を計算する。eの値を基準に確立でランダムな手、または最も行動価値の高い手を打ちます。

5.機械学習の進め方（2）

調整したハイパーパラメーターについて

- eの値(デフォルト値は1.0)
  ε-greedy法の値で、上記のような行動を取る。最も行動価値の高い手を打った時、1/20000ずつeの値を減少させる。
- バッチサイズ

一回の学習内容の規模。対戦結果をdeep-networkに反映させる際により多くの学習内容があることで特定の状況にのみ対応できるnetworkになることを回避できる。

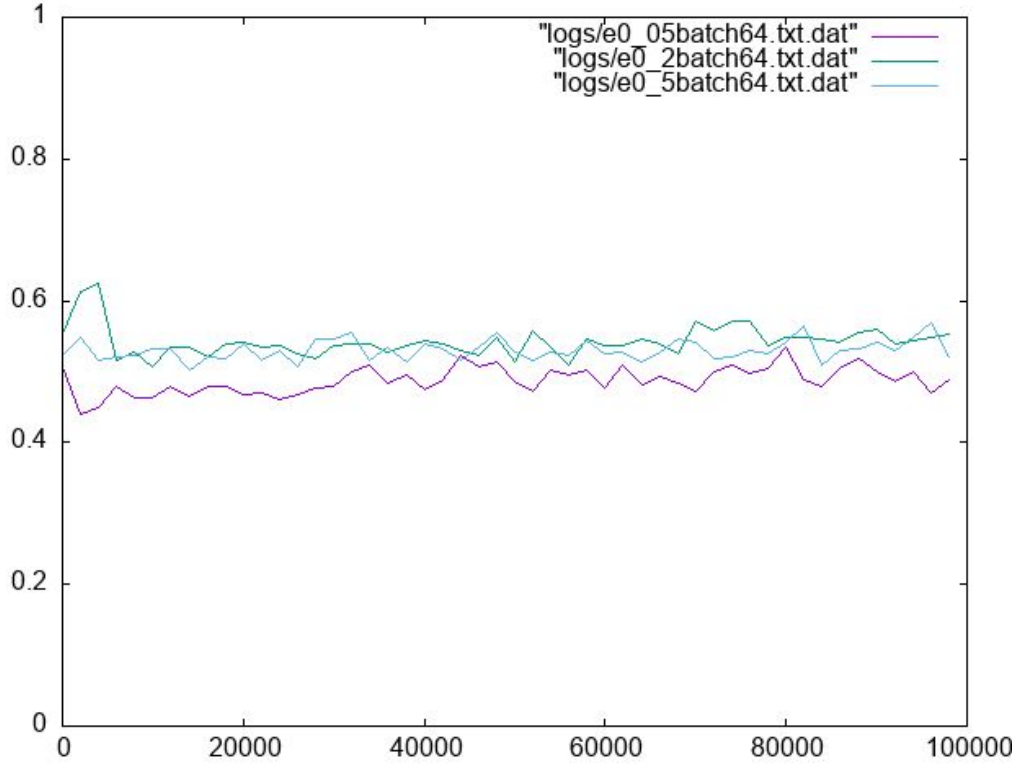
6.実験 ＜実験設計＞

DQN法を適用したプレイヤーとランダムに手を打つプレイヤーを戦わせる。

- 盤面の大きさ(4x4x4)
- バッチサイズ(64, 128, 256)とeの下限値(0.05, 0.2, 0.5)の3x3通り試し、各通り100000回試合させAIの勝率を調べる。

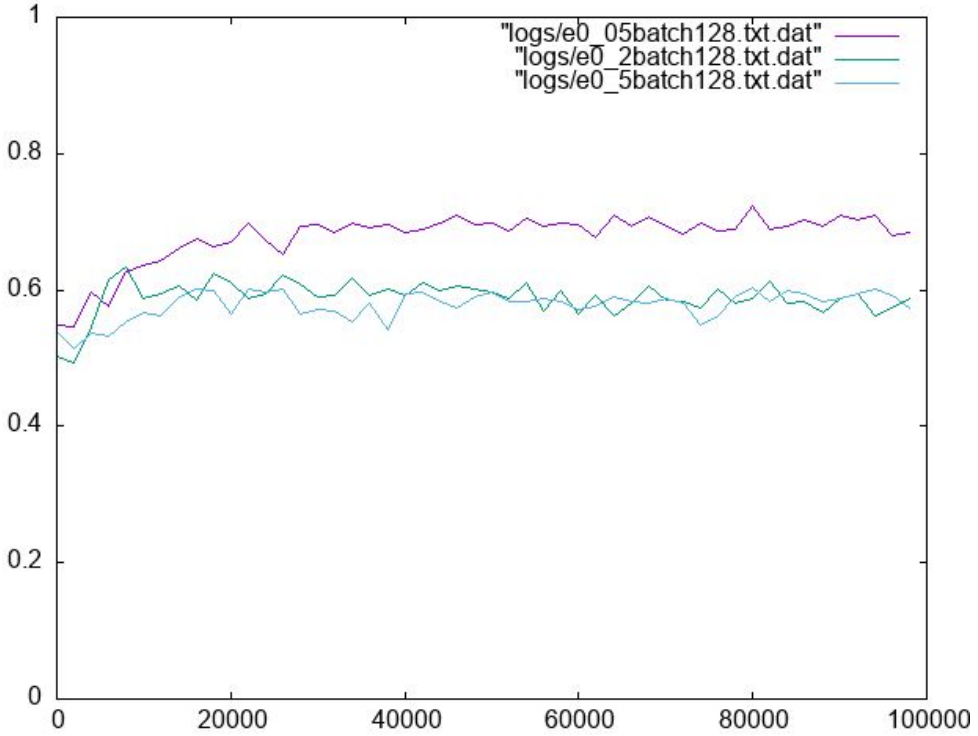
6.実験 ＜実行結果＞

- バッチサイズ = 64
- e=0.05
  - 勝率+引き分け率 0.4877
- e=0.2
  - 勝率+引き分け率 0.54253
- e=0.5
  - 勝率+引き分け率 0.53048



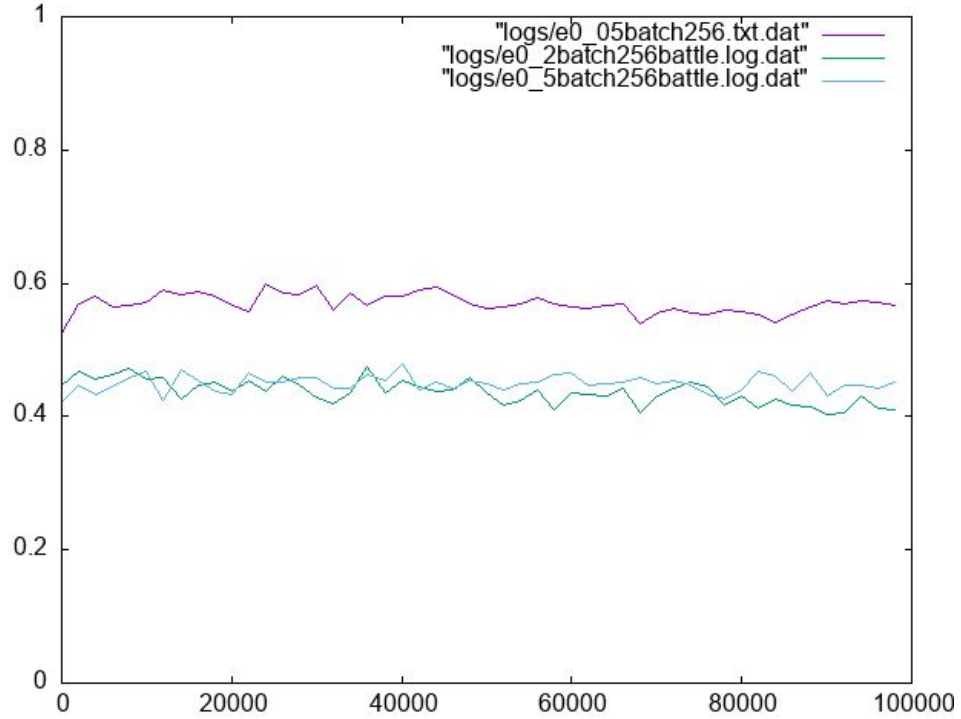
6.実験 ＜実行結果＞

- バッチサイズ = 128
- e=0.05
  - 勝率+引き分け率 0.67799
- e=0.2
  - 勝率+引き分け率 0.58875
- e=0.5
  - 勝率+引き分け率 0.57806



6.実験 ＜実行結果＞

- バッチサイズ = 256
- e=0.05
  - 勝率+引き分け率 0.56922
- e=0.2
  - 勝率+引き分け率 0.48052
- e=0.5
  - 勝率+引き分け率 0.49726



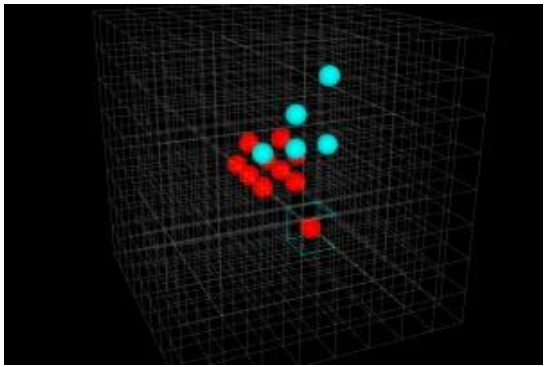
6.実験 ＜考察や自己評価＞

- eの値が低いほど勝率が上がったのは、行動価値の高い手を選択しやすいためだと考えられる。
- バッチサイズは特にeの最も低いe=0.05で比較した時、128が最も高い勝率を示した。これはバッチサイズが大き過ぎると対応できる状況が限られてしまい、小さ過ぎると十分に学習結果をネットワークに反映することが困難になるためだと考えられる。

e/バッチサイズ	64	128	256
0.5	0.53048	0.57806	0.49726
0.2	0.54253	0.58875	0.48052
0.05	0.4877	0.67799	0.56922

7.時間の都合上省いた項目、残された課題

- アルゴリズム1つのみ(DQN法)を用いて実験を行った。
- GUI化の未実施。



- アルゴリズム同士での対戦の未実施。