ISSN 2186-5647

五目並べにおける深層強化学習

日大生産工(学部)　○吉見　航　　日大生産工　山内 ゆかり

1　まえがき

近年,GoogleのDeepMind社によって開発されたコンピュータ囲碁プログラムであるAlphaGo.1)などにより,人工知能を用いたコンピュータプログラムにゲームを行わせることが注目されている.

だが,五目並べなどの状態数が膨大なものではQ学習のみではうまく学習を行うことが難しい.そこで,画像認識などでも用いられているCNN(Convolutional Neural Netowork)などと組み合わせることで,状態数を減らし,学習を行いやすくした.

本研究では五目並べにおける深層強化学習を提案し,状態数を減らすことで学習精度を高くすることを検討する.

2　研究背景および従来手法

五目並べの膨大な状態数を削減するために,盤面を区切ったうえでQ学習を行うことを試みた.

従来手法のQ-CPUの動きを下記に示す.

1. 自石または敵石が4つ並んでいるおり,5つ並んで置ける時にその場所に置く
2. FS×FSで区切られた盤面内で,自石のみがある列の自石の最大数を数える
3. 最大数をもとにQ値によって,区切られた範囲のうち1つを選ぶ
4. 選ばれた範囲にランダムで石を打つ
5. 勝敗がつくか(N×N)/2回1→4を繰り返し行う
6. Q値の更新を行う

Step6では勝った時に式(1),それ以外の時に式(2)を用いてQ値の更新を行う.

$$Q(s_t, a_t) = Q(s_t.a_t) + \alpha[R_t + \gamma max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \quad (1)$$

$$Q(s_t, a_t) = Q(s_t.a_t) * 0.9 \quad (2)$$

3　提案手法

従来手法の結果から,状態数を減らすことで学習は可能になるが,状態数を減らす際に盤面情報の損失を抑えなければ学習精度の向上を得られないことがあきらかである.

そのため,状態数を減らしたうえで盤面情報があまり失われないためにCNNとQ学習を組み合わせた手法を提案する.

4　実験結果および検討

従来手法では,10×10の二次元平面の盤面で,Q-CPUとStep1のみを取り入れたランダムに打つCPUと対戦させた.50000回の対戦を10回繰り返し,平均したものを図1に示した.

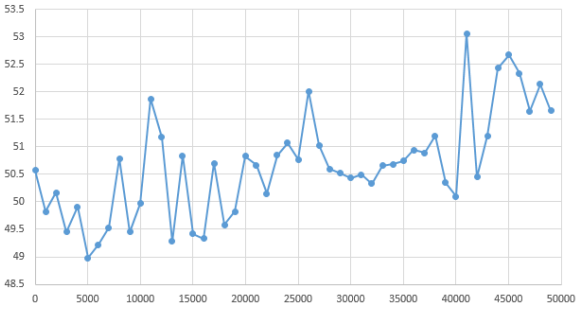

図1　1000回ごとのQ-CPUの勝率

状態数を減らすために盤面を区切ったうえでQ学習を行ったが,状態数は減らせたものの盤面情報の多くが失われてしまったので,学習はできていると考えられるが学習精度が低くなった.

それに加え,範囲選択後にランダムで手を選択しているために,Step1のみを取り入れたランダムに打つプログラム相手でも勝率はよくない結果となった.

5　まとめ

従来手法では五目並べの盤面の状態数を削減するために盤面を区切り,Q学習を行わせた.

だが,状態数の削減はできたものの盤面情報の損失が多かったのに加えて,範囲を選択後に

Deep Q-learning in Gomoku

Yoshimi WATARU　and　Yukari YAMAUCHI

ランダムで手を決めていたので,Step1のみのランダムCPUとあまり変わらない強さとなってしまった.

そのため,提案手法では今回多かった盤面情報の損失を抑えることで,学習後期にはより多く勝ちを得ることができるのではないかと考えられる.

【参考文献】

1) Mastering the Game of Go without Human Knowledge
David Silver*, Julian Schrittwieser*, Karen Simonyan*, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, Demis Hassabis.