対戦型 2048 における評価関数の強化学習と性能の比較に関する研究

1255101　小 田　駿 斗　【 高度プログラミング研究室 】

Reinforcement Learning of Evaluation Functions for Two-player 2048 Game and Their Performance Comparison

1255101　ODA, Hayato　【 High-Lebel Programming Lab. 】

1　はじめに

「対戦型 2048」は，寺田 [1] によって提案された，確率的 1 人ゲーム 2048 の 2 人プレイゲームへの拡張である．対戦型 2048 では，2 人のプレイヤ（攻撃側と防御側）が選択できる手と目標がまったく異なる．防御側プレイヤは，通常の 2048 と同様に，盤面上のタイルを動かす方向を選択し，できるだけ得点が大きくなることを目標とする．攻撃側プレイヤは，空マスの中からタイルを置くマスを選び，できるだけ得点が小さくなることを目標とする．このような非対称なゲームは対称なゲームに比べると研究が少ない．

対戦型 2048 のプレイヤの作成についていくつかの取り組みがある．岡と松崎 [2] は N タプルネットワークに基づくプレイヤを提案した．横山と松崎 [3] は，ニューラルネットワークによる評価関数と TD 誤差学習の組み合わせにより対戦型 2048 プレイヤを作成した．しかし，これら二つの研究では異なるルールを採用しており，結果を直接比較できなかった．

そこで，著者らは岡と松崎 [2] が用いたものと同じルールのもとで横山と松崎 [3] の研究と同様のニューラルネットワークを作成し，対戦実験により性能を評価した．また，岡と松崎 [2] の研究に基づいて再学習を行った N タプルネットワークを用いて，ニューラルネットワークプレイヤと同じく攻防交互に学習したプレイヤを作成した．得られたニューラルネットワークプレイヤを相互に対戦することにより性能の評価を行った．

2　ニューラルネットワークプレイヤ

2.1　評価関数

先行研究 [3] と同様に，畳み込み層 2 層と全結合層 3 層で構成されるニューラルネットワーク構造を用いて，TD 誤差学習により重みを調整した．本研究では，まず通常の 2048 における評価関数を作成し，その後，攻撃側評価関数と防御側評価関数を交互に学習した．

先行研究 [3] において，攻撃側の学習がすぐに収束するという結果が得られていたため，攻撃側の学習を 12 時間，防御側の学習を 24 時間として，それぞれ 3 回，合計 108 時間の学習を行った．また，本研究では，学習の際に対称性の考慮の有無とランダムな手を入れる割合について複数の組合せで学習した．学習の結果，対称性を考慮し，ランダムな手を入れない評価関数が最も性能が良かった [4].

2.2　探索

岡と松崎 [2] の研究において，評価関数に minimax 探索を組み合わせることで性能が向上するという結果が得られている．そこで，得られたニューラルネットワーク評価関数に $\alpha\beta$ 探索を組み合わせた [5].

2.3　実験と考察

得られたニューラルネットワーク評価関数の中で最も性能が良かった，対称性を考慮し，ランダムな手を入れない評価関数に，探索の深さ d を最大 $d=7$ とした $\alpha\beta$ 探索を組み合わせ，各探索深さのプレイヤを総当たりの対戦実験により性能を評価した．ただし，$d=2,4,6$ では相手の評価関数を利用する．対戦回数は各 100 回ずつとした．対戦結果を図 1，図 2 に示す．図 1，図 2 から，探索を深くするとプレイヤが強くなることが分かる．また，探索深さが奇数の場合と偶数の場合を比較すると，攻撃側は偶数，防御側は奇数の方が性能が良い．このことから，攻撃側の評価関数に比べて防御側の評価関数の方が優れていることが分かった．

3　N タプルネットワークプレイヤ

3.1　評価関数

先行研究 [2] で最も学習成果が良かった条件である，8 種類の 6 タプルを使用し，ランダムな手の割合を 50 % として，N タプルネットワークの学習を行った．まず，通常の 2048 における評価関数を作成し，その後，攻撃側評価関数と防御側評価関数を交互に学習した．また，学習時間はニューラルネットワークの学習と統一するため，攻撃側の学習を 12 時間，防御側の学習を 24 時間として，それぞれ 3 回，合計 108 時間の学習を行った．

3.2　探索

ニューラルネットワークと同様に，N タプルネットワークについても，$\alpha\beta$ 探索を組み合わせ，性能の確認を行った．探索の深さ d は，ニューラルネットワークプレイヤの $d=7$ と計算時間が近くなるように最大 $d=11$ とした．

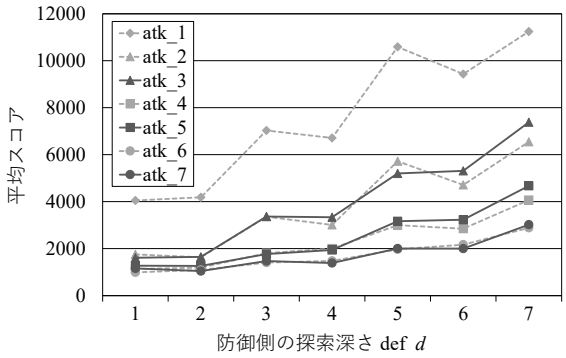

図 1　各攻撃側プレイヤの平均スコア (ニューラルネットワーク)

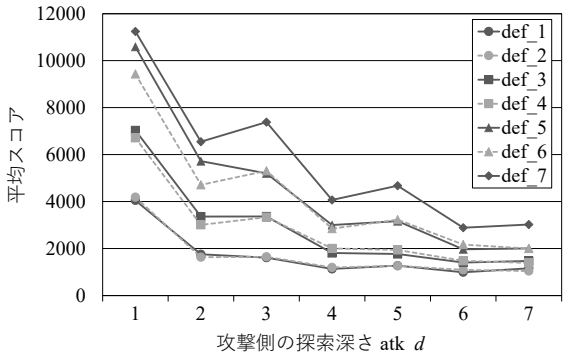

図 2　各防御側プレイヤの平均スコア (ニューラルネットワーク)

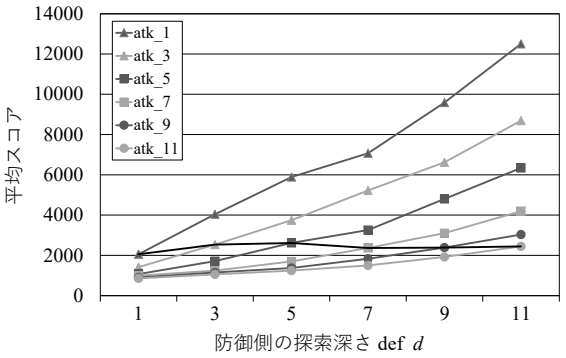

図 3　各攻撃側プレイヤの平均スコア (N タプルネットワーク)

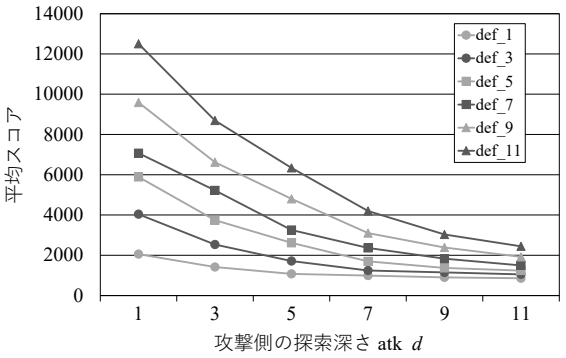

図 4　各防御側プレイヤの平均スコア (N タプルネットワーク)

3.3　実験と考察

各探索深さのプレイヤを総当たりの対戦実験により性能を評価した．対戦回数は各 100 回ずつとした．対戦結果を図 3，図 4 に示す．図 3 に引かれている黒線は，攻撃側プレイヤと防御側プレイヤで同じ探索深さの場合の結果を結んである．図 3，図 4 から，N タプルネットワークプレイヤについても探索を深くすると強くなるという結果が得られた．また，攻撃側プレイヤについて，探索を深くしていくと徐々にスコアの減少幅が下がっている．これは，ゲームの特性上，非常に小さな得点で終えることが困難であるためだと考えられる．

4　異なるプレイヤ間での対戦実験

予備実験により得られたニューラルネットワークプレイヤと N タプルネットワークプレイヤを直接対戦実験を行うことで性能を比較する．各プレイヤの探索深さ d は，着手にかかる時間が出来るだけ近くなるように，ニューラルネットワークプレイヤは $d = 7$，N タプルネットワークプレイヤは $d = 11$ として，攻防入れ替えで 100 回ずつ対戦を行った．実験の結果は，修論本体に記載する．

5　まとめ

本研究では，岡と松崎 [2] が用いたものと同じルールのもとで横山と松崎 [3] の研究と同様のニューラルネットワークを作成し，対戦実験により性能を評価した．その後，岡と松崎 [2] の研究に基づいて再学習した N タプルネットワークプレイヤを作成し，得られたニューラルネットワークプレイヤと直接対戦することにより性能の評価を行った．実験の結果，両方のプレイヤで探索を深くするごとにプレイヤが強くなること，同じ探索深さのプレイヤ同士の対戦ではスコアが 2500〜3000 に収束することが予想されることなどが結果として得られた．

参考文献

[1] 寺田　実 : 対戦型 2048，情報処理学会夏のプログラミング・シンポジウム [2015] 報告集，pp. 19–22 (2016).

[2] 岡　和人，松崎公紀 : システム的選択による N-tuple networks の “対戦型 2048” への適用，情報処理学会第 58 回プログラミング･シンポジウム，pp. 193–202 (2017).

[3] 横山智洋，松崎公紀 : ニューラルネットワークと強化学習による対戦型 2048 プレイヤの作成，情報処理学会第 62 回プログラミング・シンポジウム (2021).

[4] 小田駿斗，松崎公紀 : 攻撃側が置くタイルの数を選択できる対戦型 2048 に対するニューラルネットワークプレイヤの学習，情報処理学会第 63 回プログラミング・シンポジウム (2022).

[5] 小田駿斗，松崎公紀 : 対戦型 2048 におけるニューラルネットワークプレイヤの $\alpha\,\beta$ 探索による強化，第 27 回ゲームプログラミングワークショップ (2022).