対戦型 2048 への AlphaZero の適用と評価

1220309　岡本拓馬　【 高度プログラミング研究室 】

1　はじめに

「2048」は Gabriele Cirulli が 2014 年に公開した一人用のゲームであり，これを二人用ゲームにしたものが「対戦型 2048」である.「対戦型 2048」のプレイヤは 2 種類あり，守備側はタイルを上下左右に動かし同じタイルを揃えて高得点を目指し，攻撃側は空きマスに 2 のタイルを配置して守備プレイヤの行動を妨害する．勝敗は二人で攻撃・守備を 1 回ずつ行い，ゲーム終了時にスコアが高い方を勝利とする．

DeepMind 社が 2017 年に発表した「AlphaZero」は囲碁のゲーム AI である「AlphaGo Zero」を発展させ，囲碁以外のゲームにも対応させた汎用 AI アルゴリズムである．AlphaZero では深層学習と強化学習，モンテカルロ木探索を用いた探索を組み合わせて学習を行う．

本稿では対戦型 2048 プレイヤの学習に AlphaZero を適用し，強いプレイヤを作成できるのか実験し評価する．

2　AlphaZero の構成と適用

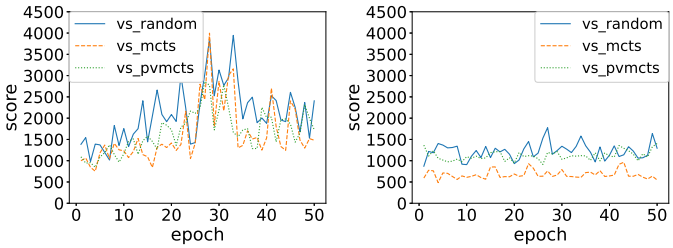
AlphaZero は初めにデュアルネットワークの構成を定義し，データ生成とパラメータ更新を繰り返すことで学習を行う．対戦型 2048 は守備側と攻撃側で盤面へのアクションが全く異なるため，それぞれでプレイヤを作成・学習し互いに対戦する．

デュアルネットワークは縦 4 マス横 4 マスの盤面について 0 から 32768 の 16 種類のタイルの位置を種類ごとに示した盤面データ $(4 \times 4 \times 16)$ を入力とし，現在の局面に応じた方策と価値を出力する．データ生成は攻撃側と守備側の最新プレイヤ同士を複数回対戦させてデュアルネットワークの学習に使用する学習データを作成する．パラメータ更新では直前のデータ生成で作成した最新の学習データを用いてデュアルネットワークの学習を行う．

本実験は守備側と攻撃側 2 つの AlphaZero プレイヤをランダムプレイヤとモンテカルロ木探索プレイヤ，そしてもう片方の AlphaZero プレイヤを対戦させることで学習したプレイヤの強さを求める．1 学習ごとに最新プレイヤは各プレイヤと 10 回ずつ対戦し，平均スコアを抽出することでモデルを評価する．

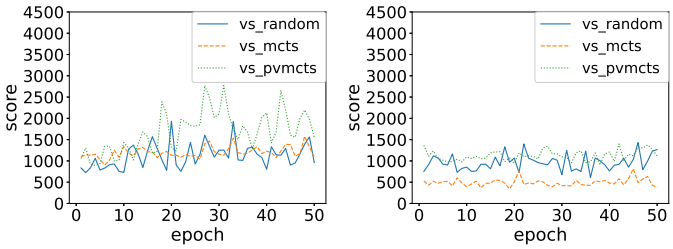
AlphaZero はデータ生成を行う際に 1 ゲーム毎に最終局面から学習プレイヤの価値を計算するが，先述の通り対戦型 2048 では盤面から勝敗を判断することは難しい．そのため本研究では勝敗条件を独自に 2 種類用意し，ゲーム終了時に予め設定しておいたタイルが存在するかで決定するタイル基準，設定したスコアをゲーム終了時に超えているかで決定するスコア基準で学習を行う．勝敗条件の初期値はタイル基準は 128，スコア基準は 1000 に設定して,「データ生成部での対戦にて半分以上で勝利」を 2 回達成した場合に基準を厳しくし，2 回達成できなかった場合は易しくする．タイル基準は指数値を 1 増減することで基準の難易度を変化させる．スコア基準は現在の基準値から守備側は達成時に +500，未達成時には −250 にし，攻撃側は達成時に −200，未達成時に +100 する．

3　結果



(a) タイル基準の結果　(b) スコア基準の結果

図 1: 守備側プレイヤの学習結果



(a) タイル基準の結果　(b) スコア基準の結果

図 2: 攻撃側プレイヤの結果

図 1 と図 2 からタイル基準とスコア基準のどちらも未学習状態から大幅に強くなることなく，特にスコア基準は守備側と攻撃側ともに学習を繰り返しても変化が全く見られなかった．タイル基準の守備プレイヤは 3 種類のプレイヤとの対戦にて最大で約 2000〜3000 点の上昇を確認できたが，先行研究で扱われてきた N タプルネットワークや CNN を用いた強化学習に比べてスコアは伸びなかった [1].

4　まとめ

本研究では，対戦型 2048 プレイヤの学習に AlphaZero を適用した．適用にあたり，データ生成時の対戦スコアから勝敗を動的に制御する方法を提案した．実験の結果，多少のスコア向上が見られたが，学習の効果は既存の他手法に比べると小さかった．

参考文献

[1] 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”, 第 62 回プログラミング・シンポジウム予稿集, 2021.