評価値の分離に基づくニューラルネットワーク 2048 プレイヤの改良

1220397　渡邊　翔太　【 高度プログラミング研究室 】

1　はじめに

「2048」は G.Cirulli が 2014 年に公開した確率的一人ゲームである．ニューラルネットワークを用いたプレイヤについては，2021 年に Matsuzaki が提案した畳み込みニューラルネットワーク (以下 CNN とする) を用いたプレイヤが最もよい結果を出している [1]．本稿ではこの CNN の構成をベースに，出力する評価値を現盤面の評価値とその差分とし，評価値については現盤面が取りうる評価値の平均で近似することとした．

2　本研究での CNN の構成

本稿で提案するネットワークは図 1 の通りである．入力には盤面の情報の 4×4 の配列 16 個により与える．これは 4×4 の盤面を 16 種類（空白, 2, 4, 8, ..., 32768）のタイルの有無による 0 か 1 の情報で表している．畳み込み層より先で 2 つに分かれ, 入力に対して 2 つの評価値を出力する．これが従来と異なる点である．

s_t の移動後かつパネル追加前の状態を s'_t とし，s'_t にランダムなパネルが追加された後の状態を s_{t+1} とする．従来のプレイヤでは以下のようにして評価値の目標値 $V'(s'_t)$ を決定し，エピソード終了時に $V'(s'_t)$ と実際の出力 $V(s'_t)$ の誤差を求め学習していた．

$$V'(s'_t) = R(s_{t+1}, a_{t+1}) + V(s'_{t+1}) \tag{1}$$

本稿で提案するプレイヤでは行動選択に用いる評価値について以下のように変更した．2 つの出力をそれぞれ $base$，adv とすると，エージェントは状態 s'_t においての行動 a_{t+1} を $V(s'_t)$ により決定する．

$$V(s'_t) = R(s_{t+1}, a_{t+1}) + base(s'_{t+1}) + adv(s'_{t+1}) \tag{2}$$

また，$base$ の目標値を平均，adv の目標値を差分とするため，それぞれに対し目標値を与える．可能な行動が k 通りのとき，s_{t+1} から行動 $a_{i,t+1}$ を行った場合の状態を $s'_{i,t+1}$ とすると目標値を以下のように表せる．

$$base'(s'_t) = \frac{1}{k}\sum_{i=1}^{k}(R(s_{t+1}, a_{i,t+1}) + base(s'_{i,t+1})) \tag{3}$$

$$adv'(s'_t) = V(s'_t) - base'(s'_t) \tag{4}$$

それぞれについて，タイル追加後に行動不能（ゲームオーバー）になった盤面については目標値を 0 とし，エピソード終了とした．プレイにおいては $V(s'_t)$ の値が最も高くなる行動 a_{t+1} を選択し, エピソード終了時に同状態における実際の出力 $base(s'_t)$，$adv(s'_t)$ と目標値 $base'(s'_t)$，$adv'(s'_t)$ のそれぞれの平均二乗誤差を損失関数とし，学習を行った．

この構成により，共通部分とその差分について効率的に学習が進むと考えた．

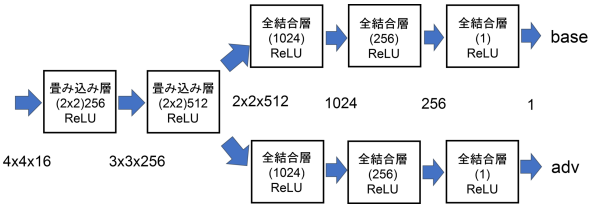

図 1: CNN の構成

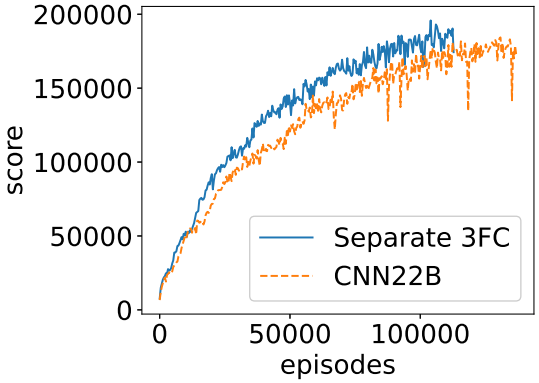

図 2: エピソード毎の獲得スコア

3　実験結果・考察

約 96 時間のエピソード毎の獲得スコアを図 2 に示す．元のプレイヤが CNN22，本研究のプレイヤが Separate 3FC である．図 2 よりエピソード毎の獲得スコアの上昇率が CNN22 よりも高いことがわかる．しかしエピソードの進行度が CNN22 と比較して低く，これはパラメータ数が約 290 万から約 526 万に増加したことが原因で計算時間が増加していることによるものと推測される．

4　まとめ

本稿では，2048 のニューラルネットワークプレイヤに対し評価値の分離といった手法での改良を提案した．結果からエピソード毎の獲得スコアは従来のものより向上しており，評価値を分離し学習させることがパフォーマンスの向上に繋がったと推測できる．しかしパラメータ数が大幅に増えたためエピソード終了までの時間が長くなった．今後の課題としてネットワークの構成の再検討，評価値の分離方法の改善などが挙げられる．

参考文献

[1] Kiminori Matsuzaki, “Developing Value Networks for Game 2048 with Reinforcement Learning”, Journal of Information Processing, Vol.29, pp.336-346, (2021)