深層強化学習を用いたポケモン対戦 AI の構築に関する研究

沈 凱弘　　指導教員：小林 邦和

1　はじめに

競技コンピュータゲームは年々大型化し，種類が増加している．一方，プレイの仕様もどんどん複雑・高度化している [1]．それを代表するのは，世界で最も人気のあるビデオゲームの 1 つのポケモンである．

近年，機械学習分野では，研究対象としてポケモン対戦に関心が集まっている．ポケモン対戦ゲームでは，機械学習にとって挑戦的な環境となるいくつかの特性が備えられている．各プレイヤーは自分のチームに関する情報と，相手のアクティブなポケモンに関する目に見える統計情報しか知らないので，部分的に観察可能である．チェスや囲碁とは異なり，環境は確率的である．ポケモンの技は区間式に基づいた命中する確率とダメージを計算する乱数を持つ．状態空間は連続的で高次元である．各プレイヤーには 6 匹のポケモンがおり，それぞれのポケモンは 18 種類のタイプから 1 つか 2 つを持つ．ポケモンごとに 4 つの技があり，それぞれにタイプもある．ポケモンごとの HP や状態、技ごとの威力と命中率の値などがある．状態空間全体を探索することは困難であるため，合理的な時間で成功する方策を実現するためには，新しい未知の状態に対してアルゴリズムを汎化する必要がある．この一連の特性は，複雑な競争ゲームを対象とし，部分的に観測可能で連続的かつ高次元の状態空間に対処できるアルゴリズムが必要である．本研究では，ポケモン対戦の人工知能 (AI) を実装し，ランキング上位プレイヤーに勝利することを目標に，最適なアプローチを探索する．

2　提案手法

本研究では，ポケモン対戦に利用する深層強化学習モデルを 2 段階で行うアプローチを提案する．第 1 段階で，人間同士が対戦したゲームログを用いて，モデルの事前学習を行う．第 2 段階で，エージェント同士の対戦データを用いて，事前学習したモデルの事後学習 (ファインチューニング) を行う．

2.1　データ

ポケモン対戦はターン毎に行動を選択する必要があるため，毎ターンの統計情報をバトルログ (replay) やリアルタイムバトルから抽出する必要がある．本研究に使用するゲームログは対戦シミュレーター Pokémon Showdown![2] の公式ページから取得できるバトルログになる．多くの研究者は Pokémon Showdown!の上で人工知能を用いたポケモン対戦エージェントを開発してきた [3, 4]．このシミュレーターは研究目的だけでなく人間プレイヤーも頻繁に利用している．その結果，シミュレーターに残る膨大なデータが機械学習に最適と考える．また，リアルタイムのターン毎の統計情報は，ライブラリ poke-env を経由し，Pokémon Showdown!から抽出されている．

ポケモン対戦をしている双方のプレイヤーは必ず自分側の情報がすべて分かる一方で，相手側の使用した技や発動した特性などは既知の情報になるが，目に見えない相手側の情報が観測できない．そのため，本研究は使用率情報で不確定な情報を補完することを提案する．replay と同様に，使用率の統計情報は Pokémon Showdown!から取得できる．データに対し，不確定情報を使用率による加重平均で表す．

2.2　ニューラルネットワーク

ゲーム中に行動の選択はプレイヤーの判断が必要な局面に対応することができるため，ポケモン対戦を POMDP[5] として扱う．選択行動の例としては，「ポケモンの交代」や「ポケモンに技を使わせる」などがある．図 1 に示したように，Actor-critic モデルを利用し，ニューラルネットワークを構築する．

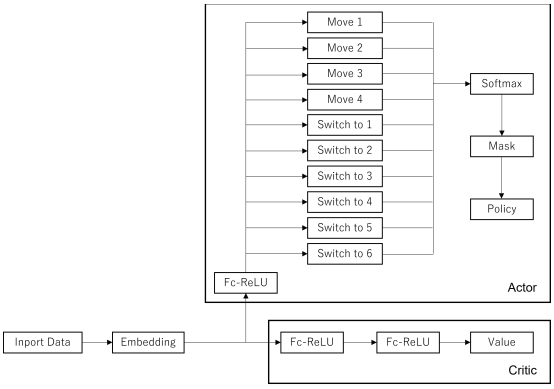

図 1　Actor-critic モデルの構成

2.3　訓練方法

訓練は事前学習と事後学習の 2 段階に分割する．事前学習は Pokémon Showdown!のランキング上位 500 位のプレイヤーが残した replay から抽出した入力データで行う．同時に，replay から抽出したプレイヤーの毎ターンの行動が教師データとなる．事前学習の訓練は，ランタムで生成するパラメータでエージェントを作成する．毎ターン終了時に，選択した行動が教師データと一致すれば正の報酬，そうでない場合は負の報酬を割り当てる．

事後学習は，事前学習のパラメータを用いて，2 体のエージェントを作成，戦わせる．試合終了時に，勝った場合は +1，負けた場合は-1 の報酬を割り当てる．学習を高速化するために，報酬を細かく与えるように設計した．補助的な報酬は，試合の毎ターンの結果に基づいて割り当てられる．

3　計算機シミュレーション

本研究は入力データの工夫により，上位プレイヤーの replay を利用し，事前学習を行うことで，短時間の学習でも良い表現の獲得が期待できる．Pokémon Showdown!上での訓練は gen8randombattle というフォーマットで行ったため，評価も同様にこのフォーマットで行う．このフォーマットは、双方プレイヤーにランダムに生成された 2 つのチームを使用させるものである．そのため，チームを手動で構築する必要がない上で，学習に必要なデータを広くカバーしている．

評価するために，gen8randombattle の上でランタムエージェント相手との勝率と人間プレイヤー相手との最高ランキング得点を取得する．ランタムエージェント相手との対戦は 1,000 回で行う．ランキング得点の計算は Pokémon Showdown!の Elo レーティングを使用する．これは，得点を 1,000 からスタート

し，勝利すると得点し，敗北すると減点になるランキング形式である．

4 結果

最終的に，事前学習では 8,066 個 replay を使用し，29 時間をかけ，約 320 回のトレーニングループ（Epoch）が完了した．また，事後学習では 43 時間で 500Epoch を完了させた．事前学習において，50Epoch 毎にプロットした損失関数と正解率の推移をそれぞれ図 2 と図 3 に示す．

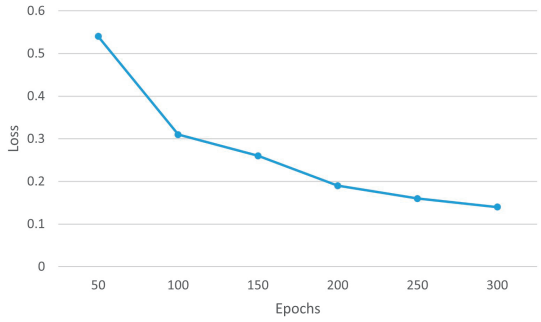

図 2 事前学習：損失関数の推移

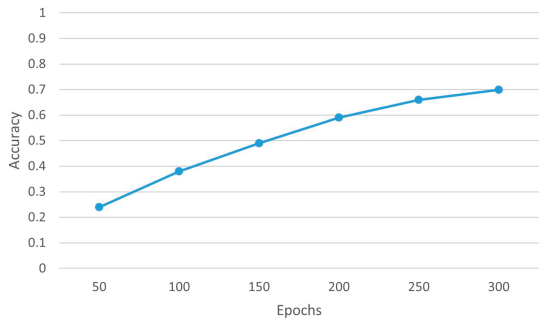

図 3 事前学習：正解率の推移

事後学習の 500 Epoch の累積報酬を図 4 に示す．このような結果は，エージェントがの勝率は平均に 50% であることを示している．これは，毎 Epoch に双方エージェントが同じ θ を使用しているためである．

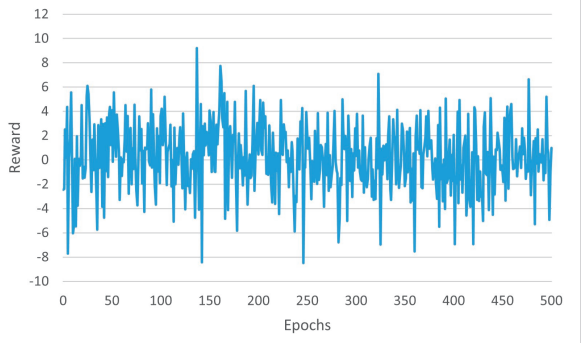

図 4 事後学習：累積報酬

1,000 回のランタムエージェントとの対戦と，200 回の人間プレイヤーとの対戦結果を表 1 に示す．ここで行った人間プレイヤーとの対戦は，Pokémon Showdown!でマッチしたランク戦のため，相手は複数人いる．また，仕様上ランクの得点が近い人間しかマッチしないため，1,000 点からスタートし，相手の実力は強くなる．最終的に，200 回の人間プレイヤーと対戦した Elo レーティングは，1543 点である．

表 1 gen8randombattle での勝敗数

Opponent	Wins	Losses
random	961	39
human player	108	92

本研究で提案するアプローチにより，ランタムエージェントに対し，96.1% の勝率を得た．また，人間プレイヤーに対し，1543 点の Elo レーティングは約 68% のプレイヤーを超えることを示している．少ない学習回数であるが，ある程度満足な結果が得られた．しかし，Huang らの研究 [3] では，6 日をかけ，ランタムエージェント対して，99.5% の勝率を出している．また，Glicko-1 レーティングで計算したランキング得点は 72% のプレイヤーを超えている．訓練にかける時間を含めて比較すれば，少し性能が劣るが，本手法は効率的なアプローチと捉えることができる．

5 まとめ

ポケモン対戦に利用する深層強化学習モデルを 2 段階で行うアプローチを提案した．Actor-critic モデルで，バトルログを使用した事前学習を加えることで，より効率なアプローチとなっている．また，より正確なデータを取得するために，事前学習と事後学習の入力データを使用率データで補完し，不確定要素のベクトル化手法を提案した．計算機シミュレーションの結果より，提案手法を用いて，訓練時間を大幅に削減しても，強力なエージェントを生成できることを確認した．これは，複雑な競争ゲームを対象とする機械学習モデルがより効率な学習方法があることを示している．

学習を 2 段階にすることにより，学習効率の上昇が見られたが，バトルログである replay の前処理に 1 ヶ月を要した．膨大な作業量に加え，ゲームの世帯交代や対戦ルール更新による仕様の変化もある．その結果，学習に使用するすべてのデータを再収集・加工しなければならない．ポケモン対戦だけではなく，複雑な競争ゲームを対象として，効率なデータ収集方法は，長期的な課題になる．

参考文献

[1] 小山友介: 「日本デジタルゲーム産業史」, 人文書院 (2016)
[2] Zarel: “Pokémon Showdown”, https://pokemonshowdown.com/, (2019)
[3] D. Huang and S. Lee: “A Self-Play Policy Optimization Approach to Battling Pokemon”, 2019 IEEE Conference on Games (CoG), (2019)
[4] D. Simoes , S. Reis , N. Lau and L. Reis: “Competitive Deep Reinforcement Learning over a Pokemon Battling Simulator”, 2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), (2020)
[5] S. Lee and J. Togelius: “Showdown ai competition”, 2017 IEEE Conference on Computational Intelligence and Games (CIG), pp.191–198 (2017)