2022 年度卒業研究概要

キャラクタ AI 生成における学習戦略が異なった強化学習アルゴリズムの比較

大谷 紀子 研究室
1972041　小暮 岳志

1. 背景と目的

近年, ゲームシステムの高度化･複雑化に伴い, ゲームプレイヤが直接操作しないキャラクタである Non Player Character（以下 NPC）のルールをエンジニアが詳しく記述することは困難となっている．したがって，NPC が自らゲーム内の環境を認識し，自律的に意思決定するためのキャラクタ AI の導入が進んでいる．キャラクタ AI を生成する手法の一つとして，強化学習がある．強化学習とは，行動の主体である「エージェント」が，与えられた環境下で戦略的かつ最適な行動パターンを学習する手法である．同じ機械学習に分類される教師あり学習や教師なし学習と異なり，学習データを使用せず自身の試行錯誤のみで学習するという特徴がある.強化学習の多様なアルゴリズムは，状態遷移と報酬を予測する関数の使用有無，保存された過去の経験の使用有無で分類される．高性能なキャラクタ AI を生成するためには，ゲームの環境に応じて適切な強化学習アルゴリズムを選択する必要がある．

本研究では，性能の高いキャラクタ AI の生成を目的とする．簡易的なゲームに対して異なるアルゴリズムで学習させた複数種類のキャラクタ AI を生成し, 強化学習アルゴリズムの性能を比較する．

2. キャラクタ AI の生成手法

強化学習アルゴリズムを分類する基準の一つに，オンポリシーとオフポリシーがある．ポリシーとは，次の行動を決定するための戦略である．オンポリシーのアルゴリズムでは，現在のポリシーで得た経験のみを利用して，新しいポリシーを予測する．一方，オフポリシーのアルゴリズムでは，保存された過去の経験を利用して，新しいポリシーを予測する．本研究では，オンポリシーである Proximal Policy Optimization（以下 PPO）[1]と，オフポリシーである Soft Actor-Critic（以下 SAC）[2]を使用してキャラクタ AI を生成する．PPO は，高い報酬が得られる行動を優先し，低い評価しか得られない行動を避けるように最適化する強化学習手法である．SAC は，Q 学習と深層学習を組み合わせた Deep Q-network に対し，局所最適解に陥らないように改良が施された深層強化学習の手法である．Q 学習とは，即時的に得られる報酬の価値と次の状態で得られる価値の和である Q 値を更新することにより，報酬の最大化を図る強化学習手法である．

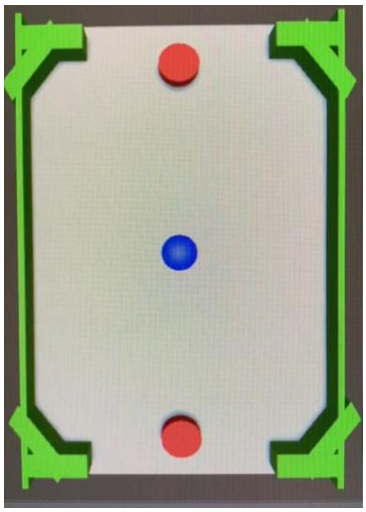

図 1：ゲーム画面

本研究では，ゲーム開発プラットフォームである Unity を使用し，エアホッケーを模した簡易的

な 1 対 1 の対戦ゲームを制作した．ゲーム画面を図 1 に示す．エアホッケーとは，プレイヤがスマッシャーと呼ばれる円盤の道具を動かしながらパックを打ち合い，相手ゴールに入れ得点を競う競技である．また，Unity で作成したシミュレータ上で強化学習を行うための機械学習ツールキットである Unity Machine Learning Agents（以下 ML-Agents）を使用した．

ML-Agents における強化学習手順を以下に示す．

① エージェントが，自身の位置・自身の移動速度・パックの位置を取得する．

② エージェントがあらかじめ定められたパターンの中から行動を選択し，実行する．

③ 行動した結果によって，報酬が与えられる．相手に勝利した場合にプラスの報酬を与え，敗北した場合にマイナスの報酬を与える．また，試合時間が長くなるにつれて報酬を増加させる．

以上の手順を繰り返し，選択された強化学習アルゴリズムに基づいて，より多くの報酬を得られる最適な行動パターンを学習する．

3. キャラクタ AI の比較

PPO と SAC を用いて 40 万ステップずつ学習させ，キャラクタ AI を生成した．累計報酬の推移を図 2 に示す．また，相対評価で実力を表すために使われる指標のひとつである ELO レーティングの推移を図 3 に示す．SAC で生成した AI は PPO で生成した AI と比較して，勝負に負けずに耐久する動きを学習した結果，試合時間による報酬を多く得た．

4. 考察

本研究で使用した環境において，PPO と SAC で生成したキャラクタ AI を比較すると，より適した行動パターンの学習スピードや相対的な強さにおいて，SAC の方が優れている結果となった．しかし，今回の環境は unity 内の限定的な場面であり，他のシチュエーションにおいて一概に同様であるとは限らない．また，エージェントは左右上下にスマッシャーを動かすのみで，戦略の幅が少なかったことも問題点として挙げられる．今後の課題として，さらに多くのシチュエーションで検証し，アルゴリズムごとの特徴を理解することが，高性能なキャラクタ AI の生成につながると考えられる．

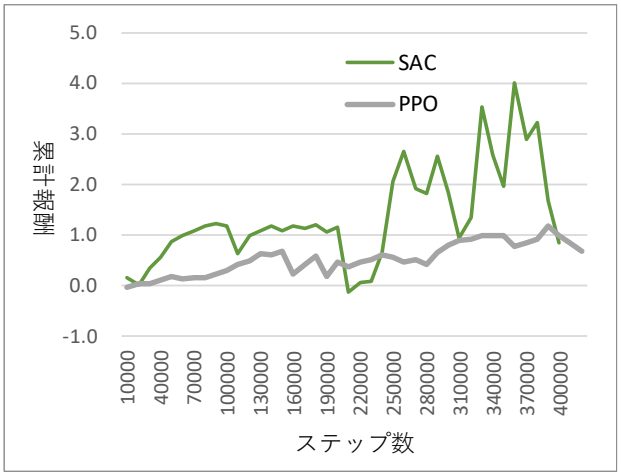

図 2：累計報酬の推移

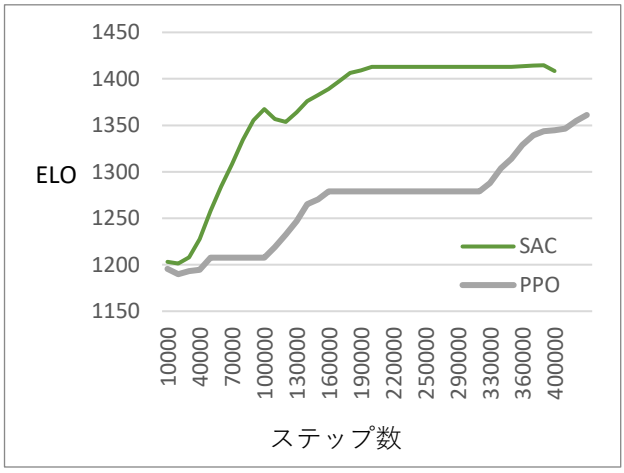

図 3：ELO レーティングの推移

参考文献

[1] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, “Proximal Policy Optimization Algorithms”, CoRR, abs/1707.06347, 2017.

[2] T. Haarnoja, A. Zhou, P. Abbee, S. Levine, “Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor”, CoRR, abs/1801.01290, 2018.