2Q-01

複数の強化学習アルゴリズムを用いたゲームAIの性能と特徴に関する考察

野替　仁太朗 † 大津　金光 †† 横田　隆史 ††
† 宇都宮大学工学部情報工学科　†† 宇都宮大学大学院地域創成科学研究科

1　はじめに

近年，ゲームのプレイヤーAIにおける進歩が著しい．例えば，2016年に囲碁のチャンピオンに勝利した「AlphaGo」がある．2018年にはその改良版であるAlphaZeroが，2019年にはMueZeroが誕生した．[1]

David Silverらの論文によると,「ミズ・パックマン」では．1面をクリアできるくらいのスコアであった．[1]

我々はどんなゲームに対しても強いゲームAIを作ることを目指している．その実現のためには，既存の強化学習アルゴリズムの特徴と限界を探ることが必要である．そのために，本稿では代表的な強化学習アルゴリズムを様々なゲームに適用して，各アルゴリズム間でのスコアを比較し，特徴と限界について考察する．

2　強化学習

強化学習とは，試行錯誤を通じてより高い報酬を得ようと試みようとする学習制御の枠組みである．報酬という情報を手がかりに学習する．状態入力に対する正しい行動出力を明示する教師は存在しない [?].

強化学習はモデルベースとモデルフリーに大別される．モデルベースは，環境から遷移モデルを学習する．エージェントが先を考えて，可能な選択肢の範囲で何が起こるかを見て行動を決定する手法である．モデルフリーは，環境から遷移モデルを学習させず，エージェントが行動をすることによって経験を蓄積し，報酬を最大化する手法である．モデルフリーの中でも，オンポリシーとオフポリシーに大別できる．オンポリシーは，現在のポリシーで得られた経験のみを利用して，次のポリシーを予測する手法である．高得点を出しにくいが，学習が安定する．一方，オフポリシーは，過去の経験を用いて，現在のポリシーを予測する手法．高得点を出しやすいが，学習が不安定になりやすい．

ここで，各アルゴリズムの簡単な説明を述べる．

ACERは，自分自身の評価を行ってからそれを更新するActor-Criticの一つである．自分自身の評価を行ってから，それを更新する手法である．十分な試行回数に達するとステップ数の残りにかかわらず学習をやめてしまう [2]．なお，文献はACERの祖先であるactor-criticについて説明したものである．確率的な行動選択ができることから採用した．

DQNは，Q学習に深層学習を組み合わせたものである．オフポリシーアルゴリズムの一種である [3].

PPOは，高い報酬が得られる行動を優先して，低い報酬しか得られない行動を避ける手法．オンポリシーアルゴリズムの一種 [4].

本論文では，オンポリシー，オフポリシーアルゴリズムで結果がどう変わるのかを調べるという目的から，ACER, DQN, PPOを用いた．

3　各アルゴリズムの特徴

実際に，ACER，DQN，PPO2 を用いて，Cartpole,Mountaincar,MsPacman,SpaceInvaders,Tennis,Roulette のゲームを評価実験してみた．平均点と分散をまとめた表を表1に示す．

表 1: 各ゲームに対するアルゴリズムごとのスコア

	ACER		DQN		PPO	
	平均	分散	平均	分散	平均	分散
cartpole	469.5	11907.9	53.2	6142.6	472.0	9417.3
Mountaincar	-200	0	-157.3	653.6	-116.9	466.1
Mspacman	210.0	0	620.1	124065.9	138.3	3784.6
Spaceinvaders	285.0	0	223.7	21635.4	269.4	269.4
Tennis	-0.159	0.279	-23.7	0.790	-23.9	0.0620
Roulette	-0.0508	1.81	-0.0971	3.97	-0.000619	0.0146

表1から，平均点については，CartpoleやMountaincar等の単純なゲームについてはPPO2が高くなっている．分散については，MsPacmanやSpaceInvadersやTennisといった複雑なゲームについてはDQNが高くなっている．よって，DQNは学習が不安定になっていることがわかる．一方，ACERについては，Mountaincar,MsPacman,Spaceinvadersの3つにおいて，学習を進めると点数が全く変動しなくなったことが確認された．ACERには目的の試行回数になったら学習をやめるという特徴をもつことに起因していると考えられる．

このことから，各アルゴリズムの特徴は以下のようになる．

ACER:指定したステップ数に達する前に学習をやめてしまうことがある．本実験では特徴をつかめなかった．

DQN:分散の値が大きくなっているので，安定性は低い．一方，平均点に関しては他の2つのアルゴリズムと同等か，それより高かったものが多かったので，最高点を出しやすい傾向がある．よって，最高点を目指すことが目的のゲームにおいて高得点を出しやすいと考えられる．

PPO2:分散の値が他の2つよりも小さいことから，比較的安定した成績が得られると考えられる．単純なゲームでの平均点が高かったので，単純なゲームにおいては比較的良い成績が出しやすいのではないか．

Consideration on the performance and features of game AI using multiple reinforcement learning algorithms
†Jintaro Nogae
††Kanemitsu Otsu,Takashi Yokota
Department of Information Science, Faculty of Engineering, Utsunomiya University (†)
Graduate School of Regional Development and Creativity, Utsunomiya University (††)

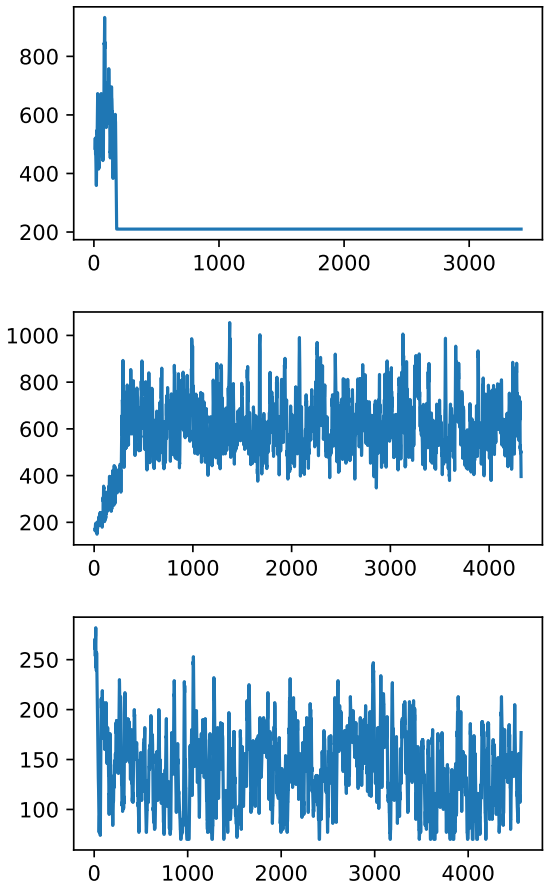

図 1: MsPacman での学習の推移

このことを踏まえると，単純なゲームに比較的強い PPO2 と，複雑なゲームについては点数が出るときもある DQN と，その他の要素をうまく組み合わせて組み合わせてどのアルゴリズムが最適なのかを学習させれば，汎用的に強いゲーム AI を実現できるのではないか．

4 強い汎用ゲーム AI の実現

強い汎用ゲーム AI を作るには，各アルゴリズムで検証した結果を考察して，成績の良かったものを選択できるようにすれば良い．最高点を目指すべきなのか，継続して高得点を出せるようにするのかはゲームによって異なる．以下に，強い汎用ゲーム AI を実現するための手順は以下のとおりである．

1. ゲームを複数の強化学習アルゴリズムで学習させる．

2. 1 で学習させたアルゴリズムを，メタレベルの強化学習アルゴリズムを使用し，高いスコアを出したアルゴリズムに対してどのアルゴリズムが最適なのかを学習させる．

5 終わりに

本稿では，汎用的に強いゲーム AI の実現に向けた下準備として，ACER,DQN,PPO2 を用いて 6 つのゲームについての比較検討を行い，それぞれのアルゴリズムの性能と特徴について考察した．その結果，単純なゲームについては PPO2，複雑なゲームについては DQN が比較的有効な手法であることがわかった．

今後の課題については，今回の結果を踏まえて，汎用的に強い独自のゲーム AI を作ることである．

謝辞

本研究は，一部日本学術振興会科学研究費補助金 (基盤研究 (C)20K11726) の援助による．

参考文献

[1] J.Schrittwieser, I.Antonoglou, T.Hubert, K.Simonyan, L.Sifre, S.Schmitt, A.Guez, E.Lockhart, D.Hassabis, T.Graepel, T.Lillicrap , D.Silver: “Mastering Atari, Go, chess and shogi by planning with a learned model, ”, Nature, vol.588, pp.604-612, 2020.

[2] A.C.Batro,R.S.Sutton,C.W.Anderson “Neuronlike adaptive elements that can solve difficult learning control problems, ” IEEE Transactions on Systems, Man, and Cybernetics vol.SMC-13 pp834-846 1983

[3] V.Mnih, K.Kavukcuoglu, D.Silver, A.A.Rusu, J.Veness, M.G.Bellemare, A.Graves, M.Riedmiller, A.K.Fidjeland, G.Ostrovski, S.Petersen, C.Beattie, A.Sadik, I. Antonoglou, H.King, D.Kumaran, D.Wierstra, S.Legg & D.Hassabis, “Human-level control through deep reinforcement learning, ” Nature vol.518 pp529-533 2015

[4] L.Engstrom, A.Ilyas, S.Santurkar, D.Tsipras, F.Janoos, L.Rudolph, A.Madry : “Implementation Matters in Deep RL: A Case Study on PPO and TRPO, ” International Conference on Learning Representations, 2020, https://openreview.net/forum?id=r1etN1rtPB.