ローグライクゲームへの深層強化学習手法の適用に関する研究

髙橋 一帆　　指導教員：小林 邦和

1 はじめに

チェスや将棋をはじめとするゲームを対象に，人工知能を開発する試みは古くから行われている．2013 年には DeepMind から Deep Q-Network(DQN) という深層強化学習手法が発表され [1]，深層強化学習による人工知能 (AI) の研究開発は加速した．囲碁，将棋，チェスなどの完全情報ゲームでは，AI が人間を打ち負かした例が報告されている．そんななかで，不完全情報ゲームの解法に研究が移ってきており，深層強化学習が対象とするゲームは，そのジャンルを広げている．今回はそのなかで，その特徴と，ゲーム自体の適度な難しさから，ローグライクというジャンルのゲーム (ローグライクゲーム) に着目した．

本研究では，ローグライクゲームの一種である Rogue-gym[2] を環境に用いた．その環境に対し，A2C，ACER，PPO の 3 種の深層強化学習手法を適用した．また，最適化する方策も，Cnn と Lstm を組み合わせた，3 種を用意した．それらの比較実験に加えて，2 種類の状態表現を用いた比較と，メタ情報の有無による比較を行った．その結果から，ローグライクゲームにおいて，上記の要素がどのように作用するかまとめる．

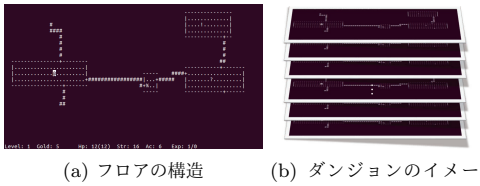
2 ローグライクゲーム

2.1 特徴

研究対象としてのローグライクゲームの特徴を述べる．囲碁やチェスは，ゲームの状況を表す全ての状態が確認できる完全情報ゲームである．これに対しローグライクゲームは，観測できる範囲が限られている．このようなゲームを不完全情報ゲームと呼ぶ．不完全情報ゲームは部分観測マルコフ決定過程 [3] を満たす典型的な問題であり，多くの探索を必要とすることから難しく挑戦的な問題と言われている．また，ローグライクゲームは行動に階層性があるという特徴を持ち，移動などの行動が，上位の行動の一部となる．ここで，報酬であるゴールドを手に入れるという行動は，移動など下位の行動の集合であるが，報酬はゴールドを手に入れたステップでのみ得られる．これは，報酬の遅延であり，行動に階層性をもつゲームでより顕著に表れる．

2.2 構造

本研究で用いるローグライクゲームは，探索範囲として，階層構造を持つダンジョンを持つ．階層構造の 1 階層は，部屋 4 つとそれらをつなぐ通路で構成されたフロアと呼ぶものであり，フロアの 1 箇所を別フロアへ繋がる階段とし，複数のフロアを接続したものを，ダンジョンと呼ぶ．フロアの構造と，ダンジョンのイメージを図 2 に示す．



(a) フロアの構造　　(b) ダンジョンのイメージ

図 1: ローグライクゲームの構造

環境は，ニューラルネットワーク (NN) への入力として，プレイヤーが観測した範囲のフロアの状態を出力する．NN は，環境への出力として，プレイヤーキャラクターを操作するコマンドを出力する．

3 深層強化学習手法

3.1 Advantage Actor-Critic (A2C)

A2C は，Actor-Critic を利用した強化学習手法の一つである．同じく Actor-Critic を利用した強化学習手法の，Asynchronous Advantage Actor-Critic (A3C)[4] の Asynchronous(非同期) を別の考え方に置き換えた手法である．複数スレッドで並列して実行する．その後，全スレッドが一定ステップ終わるまで待ち，それぞれの環境で得られた経験から勾配を計算，同期する．ざまざまなアルゴリズムが，A3C を元に開発された．

3.2 Actor Critic with experience replay (ACER)

ACER[5] は，A3C を元に，DQN で用いられる経験再生 (Experience Replay) を導入したアルゴリズムである．Experience Replay とは，各ステップで，行動とその結果を経験として保存しておき，NN の更新時にランダムにサンプリングした経験を用いる手法である．時間的相関のある似たような経験で複数回更新を行うことによって，局所解に陥ってしまうという問題を解決する．Experience Replay を用いるため，A3C をオフポリシー型に書き換えている．

3.3 Proximal Policy Optimization (PPO)

PPO[6] は，方策が出力する確率が，更新によって大きく変化することにより生じる問題を解決する手法である．同じ考えで提案された Trust Region Policy Optimization (TRPO)[7] が，KL ダイバージェンスの大きさで，大きな変化を制限するのに対し，更新前の確率と更新後の確率の比を用いて変化を制限する．これにより，TRPO が抱えていた，実装が複雑，ドロップアウト手法が使えない，Actor-Critic に適用できない，という 3 つの問題を解決した．Atari のゲーム [8] に対しては，ACER と同等のパフォーマンスを示す．その実装のしやすさと優れたパフォーマンスから，OpenAI のデフォルトアルゴリズムとなっている．

4 方策

4.1 方策 1(Cnn)

方策 1 は，3 層の畳み込み層と，1 層の全結合層で構成される．表 1,2 にその構成とパラメータを示す．

表 1: 方策 1 の構成とパラメータ (1 3 層目)

順	層の種類	フィルタ数	フィルタサイズ	ストライド	活性化関数
1	畳み込み	32	4 * 4	1	relu
2	畳み込み	64	4 * 4	1	relu
3	畳み込み	64	3 * 3	1	relu

表 2: 方策 1 の構成とパラメータ (4 層目)

順	層の種類	ノード数	活性化関数
4	全結合	512	relu

4.2 方策 2(CnnLstm)

方策 2 は，方策 1 の畳み込み層と全結合層の間に， Long short-term memory (LSTM) を組み込んだ構造を持つ．LSTM

は時系列データに対する構造の一種である．これを組み込むことで，時間的相関を用いた学習を期待する．

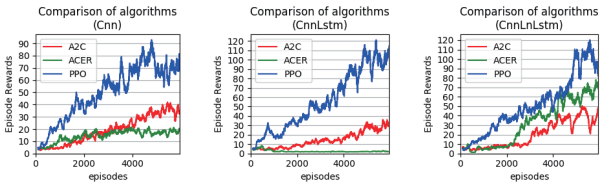
4.3 方策 3(CnnLnLstm)

方策 3 は，方策 2 の LSTM に Layer Nomalization を組み込んだ構造を持つ．

5 計算機シミュレーション

5.1 シミュレーション (アルゴリズムの比較)

用意したローグライクゲームに，A2C,ACER,PPO，3 種類のアルゴリズムを適用し，その振る舞いを検証した．各アルゴリズムでの学習結果を図 2 に示す．



(a) Cnn (b) CnnLstm (c) CnnLnLstm

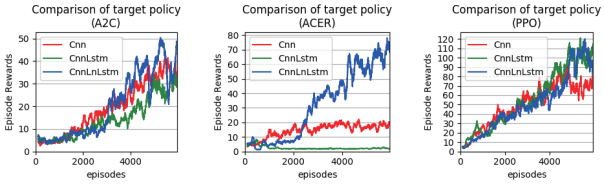
図 2: 各アルゴリズムでの学習結果

5.1.1 考察

全ての方策において，PPO がその他の手法に比べ獲得報酬が高い．このことから，2.1 で述べたローグライクゲームのような特徴を持つ環境に対しては，PPO が有効な手法であると考えられる．PPO と ACER に関して，PPO の論文 [6] で Atari のゲームに対して，ACER と近いパフォーマンスを示すとあるのに対して，本研究では差が認めらた．このことから，ローグライクゲームが，深層強化学習手法を適用する問題として，Atari のゲームと違う特徴を持つということがわかる．

5.2 シミュレーション (方策の比較)

異なる方策を用いた時，どのような振る舞いを見せるか検証する．各アルゴリズムでの学習結果を図 3 に示す．



(a) A2C (b) ACER (c) PPO

図 3: 各方策での学習結果

5.2.1 考察

PPO を用いた際の学習結果 3c から，PPO がどの方策でも同程度の獲得報酬の推移を見せており，方策による影響が少ないことがわかる．A2C も全ての方策で学習が進んでいる．一方，ACER においては，3 種の方策で異なる学習結果を示しており，方策による影響が大きいことがわかる．A2C との手法の差から，オフポリシーである点か，Experience Replay を使っている点が原因であると考えられるが，A2C の複数スレッドによる同期が，Experience Replay と似た効果を持っていることを考えると，オフポリシーであることが，方策による影響の大きさに関わっていると考えられる．

5.3 シミュレーション (状態表現の比較)

環境からの出力である状態の表現方法が，各アルゴリズムの学習結果にどのような影響を及ぼすのか検証する．2 種類の状態表現を用いて比較を行った．一方は，ダンジョンを構成する各文字を，0 1 の値に置き換え，1 チャンネルの出力とした．これをグレイスケールの出力と呼ぶ．もう一方は，各文字を 1 で表し，それ以外を 0 として，文字の種類数のチャンネルをもった出力とした．これをワンホットの出力と呼ぶ．それぞれの出力での学習結果を図 4 に示す．

5.3.1 考察

グレイスケールの場合，CNN のひとつ目の階層において，フィルターがかかる範囲には常に 0 以外の値が入っている．それに対しワンホットでは，フィルターがかかる範囲の多くの部分を 0 が占める．このことから，ワンホットの方が 1 マスだけ存在するオブジェクト (プレイヤーや階段) を認識しやすかったと考えられ，これが学習結果にあらわれていると考える．

5.4 シミュレーション (メタ情報)

環境からの出力にメタ情報を加え，各アルゴリズムの学習結果にどのような影響を及ぼすのか検証する．

メタ情報としては，プレイヤーキャラクターの足跡を用いる．プレイヤーキャラクターが通過したマスを 1，それ以外を 0 と表し，環境からの出力に追加する．図 5 にメタ情報を与えた場合の学習結果と，与えなかった場合の学習結果を示す．

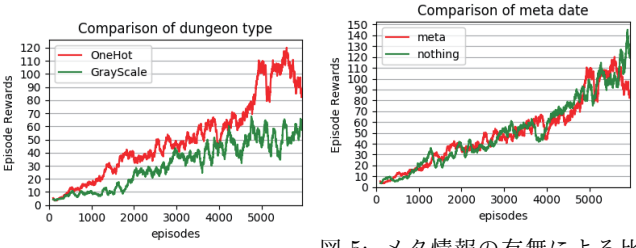


図 4: 環境からの出力の比較

図 5: メタ情報の有無による比較

5.4.1 考察

今回の比較では，方策に CnnLnLstm を用いたことから，結果に差が認められなかったと考えられる．

6 まとめと今後の展望

本研究では，ローグライクゲームの一種である Rogue-gym に，3 種の深層学習手法，3 種の方策，2 種の状態表現，メタ情報の有無で比較を行った．ローグライクゲームの Atari との差，ローグライクゲームにおける各手法，方策の性質が確認できた．今後の展望には，方策に用いた CNN の可視化が挙げられる．

参考文献

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei Rusu, Joel Veness, Marc Bellemare, Alex Graves, Martin Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, Vol. 518, pp. 529–33, 02 2015.
[2] 裕司金川, 知適金子. ローグライクゲームによる強化学習ベンチマーク環境 rogue-gym の提案. ゲームプログラミングワークショップ 2018 論文集, 第 2018 巻, pp. 120–127, nov 2018.
[3] L. P. KAELBLING. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, Vol. 101, No. 1, pp. 99–134, 1998.
[4] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *CoRR*, Vol. abs/1602.01783, , 2016.
[5] Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. *CoRR*, Vol. abs/1611.01224, , 2016.
[6] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, Vol. abs/1707.06347, , 2017.
[7] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *CoRR*, Vol. abs/1502.05477, , 2015.
[8] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *CoRR*, Vol. abs/1207.4708, , 2012.