CF-005

サブゴールによる内発的報酬を用いたモデルベース深層強化学習の考察
A Study on Model-based Deep Reinforcement Learning with Intrinsic Subgoal Reward

丸山 元輝† 遠藤 聡志† 山田 孝治†
Motoki Maruyama Satoshi Endo Koji Yamada

1 はじめに

ゲームドメインやロボティクスで近年大きな成功を収めている深層強化学習では，大きく二つのアプローチが存在する．まず一つ目がモデルの最適化である．このアプローチでは並列でエージェントを学習させる手法 [1][2] や，環境モデルを学習した上でモデルベースな手法を適応する手法 [3] が SOTA を達成している．しかしながら並列なエージェントを増やすほど，モデルベースな手法では深い先読みを行うほど膨大な計算量がかかるため，再現実験が難しい点や実環境に応用し難いという点が挙げられる．もう一つのアプローチとして，報酬関数を定義して最適化する方法がある．強化学習は適切な報酬が与えられないと思うように学習が進まない．そこでエキスパートから報酬を推定する逆強化学習や，エージェント自らが報酬を生成する内発的報酬を与えるといった手法 [4] が研究されている．この二つのアプローチより，深層生成モデルである Generative Adversarial Networks[5] を用いた数手先の先読みと，サブゴールを組み合わせた行動選択手法を提案する．ここで述べるサブゴールとは，ゴールに到達するために必要な状態，またはスタートからゴールまでの間になんども到達するような中間状態のことを指す．つまり理想的なサブゴールを設定することができれば，深い先読みを行わなくてもサブゴールを発見することで効率的な行動が可能となり，報酬がスパースな環境でもより少ない試行回数で学習が収束することが見込める．本稿ではエージェントが自律的にサブゴールを推定し，学習するアルゴリズムの開発を目標とする．

2 要素技術

2.1 Deep Q-Network

一般的な強化学習はエージェントが環境から状態を受け取り，ある方策に従って行動を選択した結果次の状態と報酬を受け取る．この一連の流れを表した図を図 1 に示す．古典的な強化学習である Q 学習では状態を離散的に扱うため，例えば画像のように高次元な状態の場合，状態数が膨大になり方策が計算不可能となる問題があった．そこで Deep Q-Network (DQN) [6] は，Q 学習の方策に深層学習を用いることによって，Atari 2600 のようなビデオゲームをゲーム画面から学習させることが可能となった. 本研究ではモデルベースな手法と DQN を組み合わせて用いる．

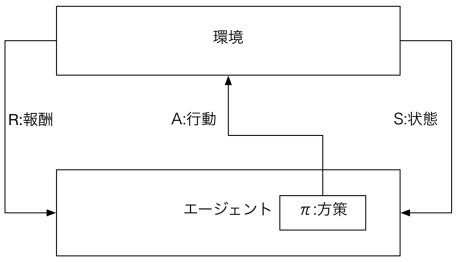

図 1 強化学習

2.2 Prioritized Experience Replay

Prioritized Experience Replay[7] とは，環境から得られる状態や報酬，エージェントの行動を保存した Replay Memory からランダムサンプリングして学習する通常の Experience Replay とは異なり，TD 誤差が高いほど優先的にサンプリングする手法である．しかしながらそのままではバイアスがかかってしまうため，TD 誤差で重みをつけたランダムサンプリングや学習時に重要度サンプリングという重みを TD 誤差にかけることによってバイアスがかかることを防いでいる．

2.3 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [5] とは，Generator と Discriminator という二つのモデルを敵対的に学習させるモデルである．基本的なアーキテクチャは図 2 に示す通り，Generator はノイズから実データに近い画像を生成しようと学習する．対して Discriminator は Generator が生成した画像に騙されないように実際の画像と識別する学習を行う．GANs はこの学習をバランスよく学習させることで高品質な画像を生成することを目的としたアルゴリズムである．Azizzadenesheli ら [8] の研究では Pix2Pix[9] をベースとしており，ドメイン変換の学習ではなく次状態を予測するための環境モデルとして学習する．

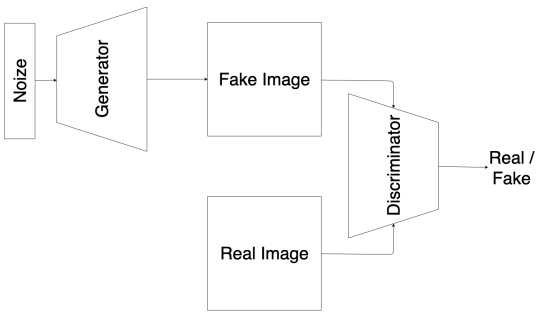

図 2 GANs のネットワーク図

† 琉球大学, University of the Ryukyus

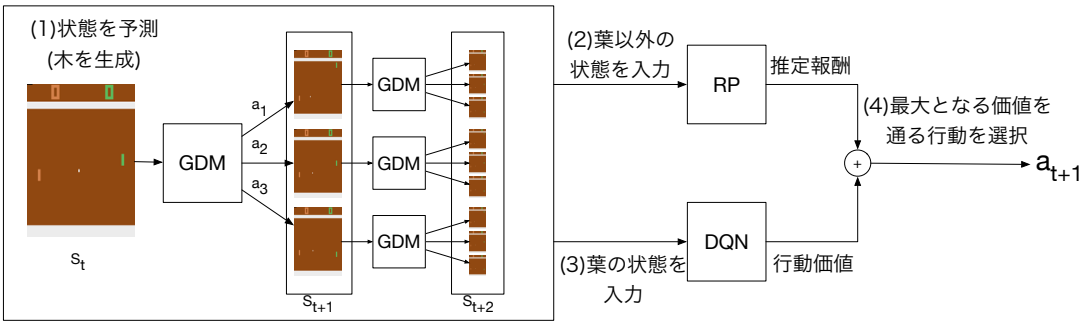

図 3　GATS の行動選択

3　Generative Adversarial Tree Search (GATS)

Generative Adversarial Tree Search (GATS) [8] とは, Monte Carlo Tree Search (MCTS) という，ボードゲーム等を効率的に解くことができる手法を Atari 2600 のようなビデオゲームで用いることを目的としたモデルベース深層強化学習手法である．一般的に MCTS を用いるためには報酬関数や状態遷移などの環境モデルが既知でなければならない．そこで Azizzadenesheli らは深層生成モデルとして注目されている GANs を用いて，環境の状態遷移を再現する Generative Dynamics Model (GDM) と報酬を推定する Reword Predictor (RP) を組み合わせて深さが有限な MTCS を実現している．GATS の行動選択手法を図 3 に示す．GDM は各行動に対してそれぞれ再帰的に次状態を生成し木を作る．次に葉以外の状態から RP が報酬を推定し，DQN は各葉の状態から最大となる行動価値を計算する．最後にそれぞれを足し合わせた中で最大の価値となる行動を選択する．

4　提案

Azizzadenesheli らの実験では，Atari 2600 のゲームのうち Pong を除くゲームで有効な結果が得られなかった．最も大きな原因として Pong 以外のゲームでは先読みの深さが足りない点が挙げられた．Pong ではボールを返すことが重要であり，数ステップの先読みでボールを返しているかどうかの状態を得ることが可能であったため攻略できた．対して Pong 以外のゲームでは，高々数ステップの先読みでは負の報酬を避ける行動のために報酬のスパース化，つまり学習の長期化を促してしまうことが問題点として挙げられた．この問題に対処するためには負の報酬を避けつつ有効そうな状態の発見，さらにその状態に対して報酬を与えることで報酬のスパース化を避けることが望まれる．そこでサブゴールを設けて，さらにそのサブゴールに対して報酬を与えることでこの問題を解決する．サブゴールと先読みを組み合わせた行動選択法を図 4 に示す．木を生成する箇所は GATS と同様だが，生成した状態とサブゴールを比較することでサブゴールが発見されればサブゴールを通るように行動，発見できなければ通常の GATS のアルゴリズムを使用する．またサブゴールを得るために以下のような流れでサブゴール探索を行う．

1. あるエピソードから状態履歴を保存
2. 状態履歴からサブゴールをサンプリング
3. エピソード終了まで行動
4. あるタイミングまで 2〜3 を繰り返す
5. 総報酬が更新されたら 1 へ戻る

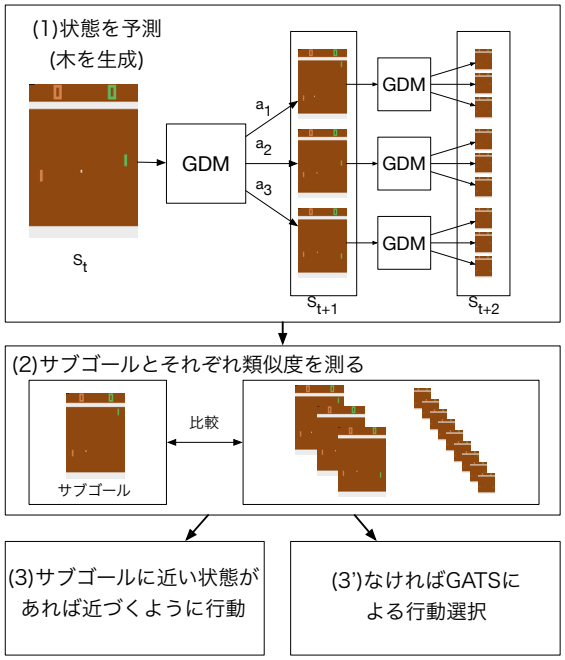

図 4　GATS とサブゴールによる行動選択

5　予備実験

まず理想的なサブゴールが取り出せている前提で学習可能なのかを調べる実験を行なった．DQN のみと先読み+DQN それぞれに対しサブゴールの有無による比較を平均報酬で評価する．理想的なサブゴールを図 5 にそれぞれ示し，またゴールした状態を最終サブゴールとして加えた．今回のサブゴールの特徴として，スタートからゴールまである程度等間隔であること，また分岐点を含んでいるかを条件として選択した．またサブゴール報酬は過学習を抑えるために，サブゴールの到達回数に応じて以下の式で計算するように設定した．

$$r_{subgoal} = \frac{1}{\sqrt{N}} \quad (\therefore N : \text{到達回数})$$

図 5　理想的なサブゴール

5.1　実験環境

今回は Pygame という 2D ゲームを作るためのライブラリを使用し，シンプルな迷路を作成した．以下に実験で使用した迷路の実験設定を示す．

- 1000 ステップでゲーム終了
- ゴールできると報酬+1
- 1 ステップ毎に報酬-0.001

以降の迷路を使用した実験ではこの設定を用いる．図 6 は実際に使用した迷路を示しており，左上の青色がプレイヤーで右下のオレンジがゴールである．またエージェントは上下左右に行動可能で，壁に当たる行動を取ったときはその場に留まる．迷路自体はランダムで生成することができるが，今回は図 6 の迷路を使用する．今回迷路を選択した理由としてはサブゴールを容易に定義し易いことが挙げられる．

図 6　迷路の全体図

5.2　結果・考察

図 7 に迷路での平均報酬の推移を示す．図 7 より，DQN では学習が進んでいないことが分かる．この理由として，報酬がスパースな環境だと行動価値関数を更新するための報酬が得られにくいため，学習が困難になってしまうことが挙げられる．先読みに関してはサブゴールの有無に関わらず学習ができており，さらにサブゴールが存在する場合平均報酬の上昇が早いことが示された．しかしながら一度平均報酬が上昇した後下降する傾向が見られる．この原因として，ゴールができるようになるにつれて行動価値関数が更新されるが，ゴール付近の状態よりもスタート地点付近の状態が圧倒的に多く Replay Memory に保存されているため，不完全な行動価値関数でスタート地点付近の行動価値が更新された結果，スタート付近とゴール付近の行動価値が逆転したために起こったと考えられる．実際に平均報酬が下がった後，途中までゴールに向かう行動をとった後に逆走する現象が確認された．

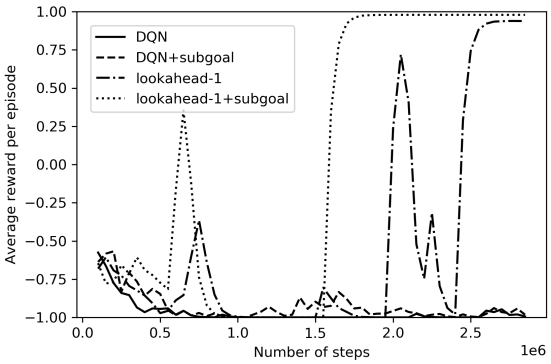

図 7　迷路での理想的なサブゴールの実験における平均報酬の推移

6　実験

予備実験の結果より，理想的なサブゴールが取得できていれば学習に大きく貢献することが確かめられた．次にエージェントがサブゴールを自律的に探索した場合ではどのような結果が現れるかを調べるための実験を行なった．また比較対象として先読みのみと先読み+理想的なサブゴールで平均報酬の比較を行なった．環境は予備実験で使用した迷路をそのまま使用する．

6.1　実験概要

サブゴールのサンプリング方法に以下に示す到達率を利用したサンプリングを使用し，Perseptual hash で類似度の計算を行なった．

$$\text{到達率} = \frac{\text{サブゴール到達回数}}{\text{サブゴール\&ゴール到達回数}}$$

今回サブゴールを探索するための状態履歴は総報酬が高いエピソードから取り出した．したがって総報酬が更新されれば今まで計算された到達率もリセットされる．また 1 エピソード毎に前回の総報酬より今回の総報酬が低い場合，再計算された到達率からサブゴールの更新を行う．今回取り出されたサブゴールの中に同じ状態が含まれる可能性があるため，その場合は重複したサブゴールを削除している．したがって取り出されるサブゴール数は，重複する事も考慮した $\max(\text{状態履歴の総数} \times 0.05, 7)$ で計算される．

6.2　結果・考察

図 8 に平均報酬の推移を示す．エージェントによるサブゴールの探索での実験結果としては，先読みのみと理想的なサブゴールでの結果と比べて平均報酬の上昇がもっとも早い．この原因として初期段階で有効なサブゴールが取り出されており，また三つの理想的なサブゴールより多くのサブゴールが取り出されていることが考えられる．一方で学習の後半に平均報酬が不安定な箇所が見られる．この現象に関して実験設定より大きく分けて二つの原因が考えられる．一つ目に総報酬の最高値の更新による状態履歴の入れ替えである．状態履歴の更新時には到達率を引き継ぐことができないため，有効なサブゴールを取り出すための重みが初期化されサンプリ

ングが不安定になった．またその際にサブゴール報酬も初期値に戻り，再度大きな報酬を渡されたことも原因として考えられる．二つ目に有効なサブゴールの固定化が不安定であることが挙げられる．今回 1 エピソード毎に前回の総報酬よりも今回の総報酬が低ければサブゴールを再取得していた．その結果平均報酬が上昇した後の微小な総報酬の変化に非常にセンシティブになり，毎回サブゴールが更新されたため平均報酬が不安定になったと考えられる．

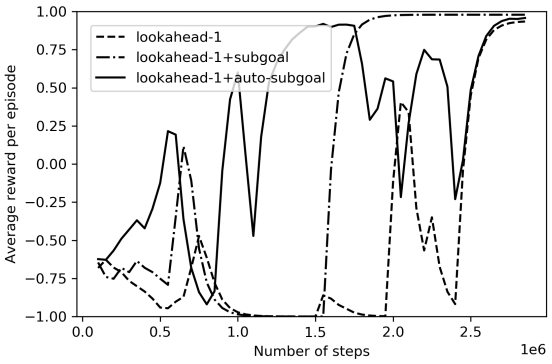

図 8　推定されたサブゴールの実験における平均報酬の推移

7　おわりに

本研究ではサブゴールの有効性とサブゴールの自律的な探索手法の有効性を示した．サブゴールを設定することで学習の収束速度に貢献し，また自律的にサブゴールを設定することでも収束することが確認できた．しかしながら懸念点として現状迷路の環境のみでしか実験を行なっていない．つまり迷路以外の環境でも学習が可能なのかを検討する必要がある．迷路ではただ一つのゴールまでのルートとゴールが存在するタスクであった．そのためサブゴールの設定が容易な環境であると言える．今後の課題として段階的にタスクを解くような環境や，タスクを解くまでに複数ルートを含むような環境で実験，考察する必要がある．

参考文献

[1] Steven Kapturowski, Georg Ostrovski, John Quan, Rémi Munos, and Will Dabney. Recurrent experience replay in distributed reinforcement learning. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[2] Dan Horgan, John Quan, David Budden, Gabriel Barth-Maron, Matteo Hessel, Hado Van Hasselt, and David Silver. Distributed prioritized experience replay. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[3] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy P. Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *ArXiv*, abs/1911.08265, 2019.

[4] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *34th International Conference on Machine Learning, ICML 2017*, 2017.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[6] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. 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*, 518(7540):529–533, February 2015.

[7] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized Experience Replay. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, nov 2015.

[8] Kamyar Azizzadenesheli, Brandon Yang, Weitang Liu, Emma Brunskill, Zachary C. Lipton, and Animashree Anandkumar. Sample-efficient deep RL with generative adversarial tree search. *CoRR*, abs/1806.05780, 2018.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.