F-003

GAN を用いたモデルベースな深層強化学習に関する考察
A Study of Model-Based Deep Reinforcement Learning Using GAN

丸山 元輝† 遠藤 聡志‡ 山田 孝治‡ 當間 愛晃‡ 赤嶺 有平‡
Motoki Maruyama Satoshi Endo Koji Yamada Naruaki Toma Yuhei Akamine

1. はじめに

近年，ゲームドメインにおいてモデルフリーな深層強化学習が人間を超える性能を見せてきた[1]．一方で，深層生成モデルの発展により，モデルベースな手法が研究され始めている[2]．モデルフリーな手法では学習に膨大な時間がかかるため，モデルベースな手法と組み合わせて学習時間を短縮させることが望ましい．本研究では，高精度な画像を生成できることで注目されている Generative Adversarial Networks[3] を用いた状態予測と，CNN による報酬予測を組み合わせたモデルベースな深層強化学習の有効性の検証，考察を行なう．また，強化学習では一般的に報酬を最大化し，かつゲームをクリアするために学習を進めるが，その途中の状態遷移はあまり考慮されない．そこで，ゴールに向かうまでに何度も通るような状態をサブゴールと定義し，予測した状態とサブゴールを比較することで，最短経路を通り，高速に学習を進める手法を提案する．

2. 関連研究

2.1 Deep Q Network (DQN)

Deep Q Network (DQN) [1] は，深層学習と古典的な強化学習である Q 学習を組み合わせた手法である．深層学習で行動価値関数を近似させることによって，ゲーム画面を入力として学習させることが可能になった．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．

2.2 Generative Adversarial Tree Search (GATS)

Generative Adversarial Tree Search (GATS)[2] とは，AlphaGo[4]のように深層強化学習と Monte Carlo Tree Search (MCTS)を組み合わせたモデルベースな深層強化学習を目的としたアルゴリズムである．実際に MCTS を用いるためには環境のモデルが既知である必要がある．そこで，画像生成で注目されている Generative Adversarial Networks (GAN) を用いて環境のモデルを再現する 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 を実現する．実際のアルゴリズムを図 1 に示す．以下のようなプロセスを行う．

(1) GDM が決められた深さ分次の状態を予測
(2) 予測された状態を利用して RP が報酬を推定
(3) DQN が葉の状態から行動価値を出力
(4) 報酬と行動価値に割引係数をかけて足し合わせ，最大となるパスを通るような行動を選択

これを行動選択毎に行うことで，最適な行動選択を可能とし，DQN のみと比べて学習速度の向上が確認されている．

3. 提案手法

GATS アルゴリズムの一つの要素である，GDM の先読みを利用して，サブゴールと組み合わせた手法を提案する．ここではゴールにたどり着くまでに何度も通るような状態をサブゴールと定義する．以下にその基本的なアルゴリズムの概要を示す．

1. サブゴールの決定
2. GDM で次の状態を予測
3. 予測した状態とサブゴールを比較
4. サブゴールが存在する場合，それを通る行動を選択

このように行動毎に先読みとサブゴールの探索を行うことで，より最適な行動選択を行うことを期待する．

4. 実験

4.1 実験概要

実験環境に Atari 2600 の一つである Pong を使用する．先読みの実験では，予測画像の妥当性を視覚的に確認し，DQN と先読みの平均報酬の比較を行う．サブゴールの実験では，ある程度学習した DQN を使用して，ゲームをプレイさせた中で総報酬が高いエピソードを選び，10 ステップごとに状態を取り出して，それをサブゴールに設定した．また，サブゴールとの比較方法は Perceptual Hash を使用しており，サブゴールとの類似度を GATS の価値に加えることで，先読みも考慮したサブゴール選択を行う．

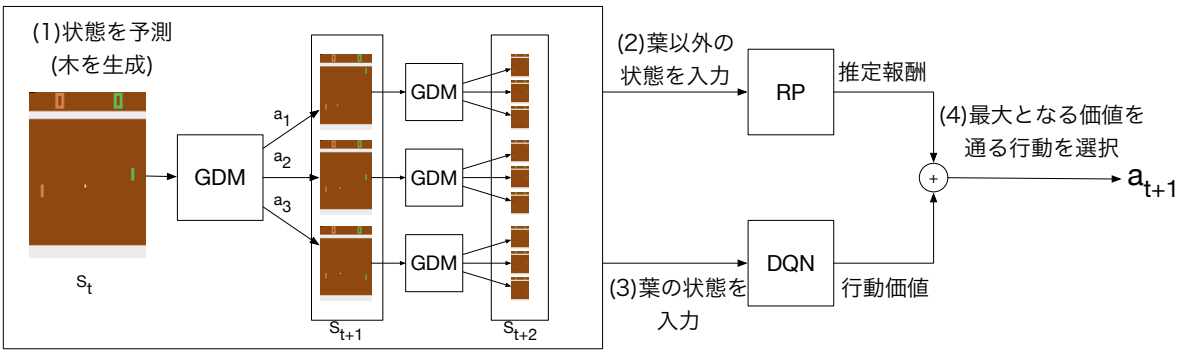

図 1. 行動空間が 3 で深さ 2 の場合での GATS

† 琉球大学大学院理工学研究科情報工学専攻，Graduate School of Engineering and Science, University of the Ryukyus
‡ 琉球大学工学部工学科知能情報コース，Computer Science and Intelligent Systems, University of the Ryukyus

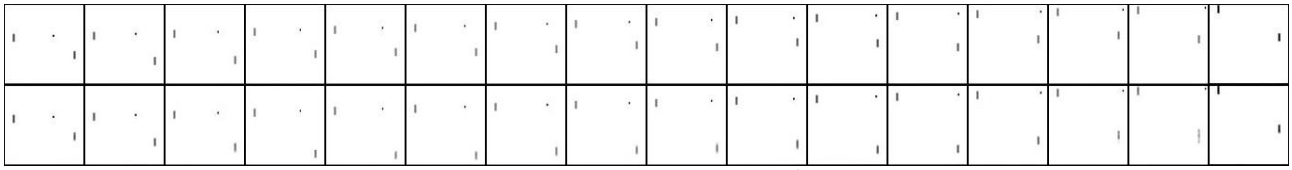

図 2. 実際に観測された状態（上）と GDM が予測した状態（下）

4.2 実験結果

4.2.1 先読みの実験

図 2 は学習初期で 16 ステップ分予測した状態を表しており，わかりやすいように白黒反転したものである． この図では先読みの深さ 1 で学習を行った GDM に，予測した状態を再帰的に入力して生成した．それにもかかわらず予測された状態の精度は高い．紙面の都合上割愛するが，さらに学習を進めていくと，ボールの反射やパドルの動きまでもほぼ再現できていることが確認できている．このことから GDM による状態の予測精度は高水準であり，モデルベースな手法で使用可能であるといえる．次に DQN と先読みの平均報酬を図 3 に示す．

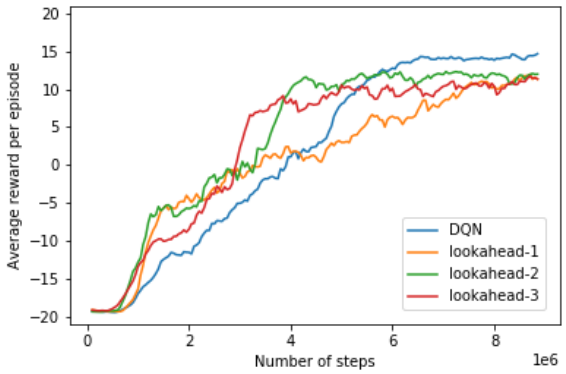

図 3. 先読みによるそれぞれの平均報酬

図 3 は DQN と lookahead-1～3 の平均報酬の推移を表している．学習初期では全ての先読みが DQN よりも平均報酬が高い．しかしながら lookahead-1 は 4M ステップあたりから平均報酬が DQN よりも低くなっている．また lookahead-2，3 も収束自体は早いが最終的な平均報酬は DQN より低い．lookahead-1 が 4M ステップ以降，DQN より平均報酬が低い理由として，学習初期では先読みのおかげでボールを跳ね返すことができているが，負の報酬が得られにくくなるため，行動価値の更新があまり行われなくなる．つまりラリーを続けることができるが，決定打を打つことができなくなるためにこのような結果になったと推測される．lookahead-2，3 の平均報酬の収束が早くなる原因は，先読みが深いほどボールを打ち返すことが間に合わないことが少なくなると考えられる．一方で，点を入れることができる最適な打ち返し方を学習する機会がより少なくなるために，最終的な平均報酬が DQN より低い結果となる．

4.2.2 サブゴールの実験

先読みのみとサブゴール＋先読みの平均報酬を図 4 に示す．学習初期では両者に変化はあまり見られない．その理由として，学習初期では学習済みの DQN を用いてサブゴールを決定しているため，サブゴールと予測画像の差が大きく，行動にあまり影響を与えないためである．また，10 ステップごとにサブゴールを取り出しているので，行動選択に有効なサブゴールがあまり取り出せていないことが予想される．今回使用した環境である Pong は，ボールを打ち返す付近がサブゴールだと考えられるため，平均報酬の大きな向上には至らなかった．しかし，3M ステップを超えたあたりからわずかに優っている箇所が見られるため，さらなる検証が必要だが，少なからずサブゴールの優位性が存在するのではないかと考えられる．

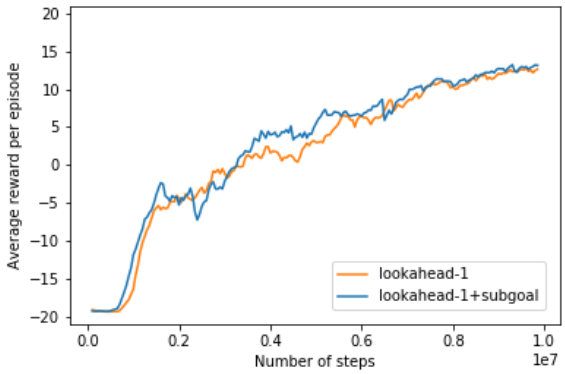

図 4. サブゴールと先読みの平均報酬

5. 終わりに

本研究では，GDM と RP による環境モデルの再現性と，それを用いたサブゴールの有効性について検証，考察を行なった．環境モデルを再現するということは現段階で概ね達成できることがわかる．しかし，Pong より複雑なゲームの場合，予測することが可能なのか，効果的に先読みの効果が発揮されるのかいまだ疑問が残る．またサブゴールについても，Perceptual Hash 以外を使用して次元圧縮する方法や，サブゴールの選び方，サブゴールに到達時に報酬を与える方法などについても議論する余地がある．

参考文献

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518 (7540): 529–533, 2015.

[2] K. Azizzadenesheli, B. Yang, W. Liu, E. Brunskill, Z. C. Lipton and A. Anandkumar. Sample-efficient deep RL with generative adversarial tree serch. *arXiv preprint arXiv: 1806.05780*, 2018.

[3] I.Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pp. 2863–2871, 2014.

[4] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529 (7587): 484–489, 2016.