7C-04

Attention Branch Network における 報酬の不確実性予測を伴う深層強化学習手法の提案

鈴木 彼方 †,‡　　尾形哲也 ‡,††

† 富士通研究所, ‡ 早稲田大学, †† 産業技術総合研究所

1 はじめに

近年，強化学習手法に神経回路モデルを用いた深層強化学習はロボット [1] やビデオゲーム [2] において高い性能を示しているが，タスクごとの環境や報酬の設計が必要で，学習が不安定という課題がある．実世界でのタスクへの深層強化学習の適用を考えると，適切な報酬を設計することは容易ではない．ビデオゲームなどでは試行により得られる報酬が明確であるが，実環境での試行を伴うケースにおいては報酬値に事前に設計できないノイズが含まれる可能性がある．例えばロボットの動作生成において，目標状態を示すセンサ情報 (例えば画像や特徴量など [3][4]) から報酬値を推定する場合を考える．最終状態をセンサ状態やそこから抽出した特徴量によって与えると，報酬値に実環境に伴う種々のノイズが含まれ，学習が不安定となる．

従来研究では，報酬の誤解釈や観測失敗がある場合におけるマルコフ決定過程を定義し，サンプリングによって真ではない報酬に対処する方法を提案している [5]．しかしながら，提案はテーブル形式の報酬に関する議論にとどまり，連続的な制御手法の提案をしていない．一方 [6] では，報酬信号の分散を減らすための推定器が提案され，サンプリングされた報酬の代わりに割引価値関数を更新する．

本研究では環境から得られる報酬に含まれる分散を直接推定するサブタスクを深層強化学習手法へ組み込み，かつ，行動を学習するネットワークに反映させることを考える．評価実験では提案手法を Open AI gym の Atari に適用し，その環境から得られる報酬にノイズを加えて性能を評価することで，有効性を検証する．

2 提案手法

本研究では強化学習手法の一種である Asynchronous Advantage Actor-Critic[7] を拡張し，報酬の分散を予測するサブタスクを解く．また，その過程で得られた特徴マップを行動出力に参照する構造を組み込む．

2.1 Asynchronous Advantage Actor-Critic (A3C)

A3C とは Asynchronous，Advantage，Actor-Critic から構成される学習手法である．複数環境における非同期学習 (Asynchronous) においては，Global Network のパラメータをコピーした worker により，異なる環境で複数のエージェントで学習したパラメータが非同期に反映される．また，数ステップ先までの報酬の推定 (Advantage) することにより，1step 先のみ推定する手法と比較して，より確からしい現状態の価値の推定誤差を利用することで安定した学習が可能となる．加えて，Actor-Clitic 法の行動と状態価値の独立した予測により，報酬の変化が行動に悪影響を与えにくくしている．

2.2 報酬の不確実性の予測

提案手法では，A3C に状態価値の学習時における予測対象の不確実性を考慮した学習 [8] を行う機構を組み込んだ．目標値 (環境から与えられる報酬) の状態にガウス分布を仮定し，その平均と分散の予測をそれぞれ出力とし，負の対数尤度を最小化するように学習を行う．状態価値 $V^{\pi}(s_t)$ の予測に関する不確実性を分散として推定するために，エピソード中のタイムステップ t における報酬 r_t の確率密度 $p(r_t|s_t,\theta)$，及び，対数尤度 L は以下により定義される．

$$p(r_t|s_t,\theta) = \frac{1}{\sqrt{2\pi\nu_t}} exp\left(-\frac{(V^{\pi}(s_t)-r_t)^2}{2\nu_t}\right) \tag{1}$$

$$L = \prod_{t=1}^{T} p(r_t|s_t,\theta) \tag{2}$$

ここで θ はモデルのパラメータである．これは出力誤差を不確実性の予測 ν で割った，重み付き予測誤差を最小化することに相当し，状態価値の学習が安定化する．

2.3 Attention Branch Network

状態価値の予測の安定化を行動予測にも反映させるため，提案手法では Attention Branch Network (ABN)[9] を用いる．状態価値と Attention map を出力する Value branch，Attention map を考慮して行動を出力する Policy branch からなるモデル [10] に，前述した報酬の不確実予測を行う Variance branch を加えたモデルを構築した．

モデルは画像を入力とし，Value branch により状態価値を，Variance branch により状態価値の分散を出力する．ここで，Value branch から出力された特徴マップ $f(s_t)$ の最大値を状態価値として扱う．加えて，$f(s_t)$ は下記の Residual 機構に基づき，Policy branch の特徴マップ $g(s_t)$ に足し合わせられる．

$$g'(s_t) = (1 + f(s_t)) * g(s_t) \tag{3}$$

Deep Reinforcement Learning with variance estimation of reward in Attention Branch Network

Kanata Suzuki (Fujitsu Laboratories LTD. and Waseda Univ.), Tetsuya Ogata (Waseda Univ. and AIST)

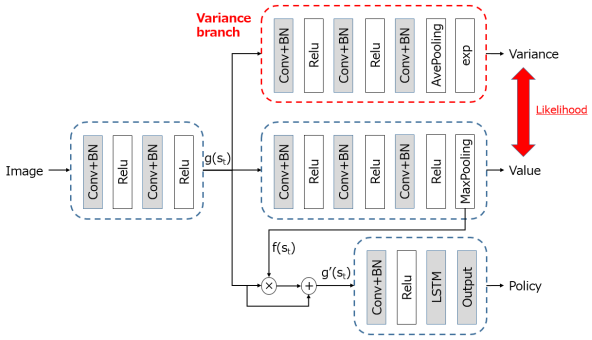

図 1: ネットワーク構造

これにより環境 s_t における状態価値を行動に反映し，かつ，元の特徴マップの消失を抑制する．最終的に上記の $g'(s_t)$ を LSTM らに入力することで行動を出力する．

3 評価実験

3.1 実験設定

評価実験では，提案手法を Open AI Gym のゲーム環境の一種である BreakOut に適用した．BreakOut ではでは操作画像をネットワークに入力し，ゲーム内の操作対象をエージェントとしてブロックを崩してスコアを獲得する．使用するネットワークの構造を図 1 に示す．入力画像は任意の領域を切り出した後に，80 × 80 のグレースケール画像に変換する．学習時には最適化手法に RMSprop を用い，学習係数を 7e-4，割引率を 0.99，worker の数を 32 とした．

また，環境から観測される報酬値にノイズが含まれる場合を想定し，大きさの異なる分散 σ^2 に従うガウス分布に基づくノイズが加えられている．実験では σ^2 はそれぞれ 0.0，0.03，0.05 とした．分散を推定する機構のない ABN-A3C[10] をベースとし，提案手法の有無による学習安定化に関する検証を行う．

3.2 実験結果と考察

ゲームの学習過程におけるスコア遷移を図 2 に示す．図中の縦軸はゲームで得られた点数を，横軸は worker の総学習回数を示す．Variance branch 構造を持つ提案手法を赤線で，ABN-A3C を青線で示す．各条件における学習結果を比較すると，分散が大きくなると学習結果のばらつきが大きくなることがわかる．そのなかで提案手法は ABN-A3C と比較して，分散の大きさに関わらず学習の収束が早く，かつ，ばらつきが小さい．特に分散が小さい条件 ($\sigma^2 = 0.0, 0.03$) において，提案手法は ABN-A3C の最もよいスコアに近い性能となった．これは提案手法が学習時に報酬の分散を予測し，行動に関するネットワークの学習が安定したと考えられる．

ただし，報酬に含まれるノイズが大きくなると ($\sigma^2 = 0.05$)，ABN-A3C と比較して良い性能であるものの，提案手法のスコアはばらつき始め収束が遅くなった．特徴マップの接続方法や，分散学習を行う係数などのパラメータ調整に改善の余地があると考えられる．

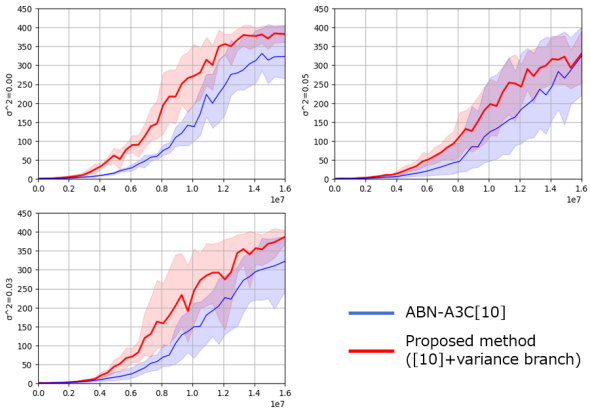

図 2: 各条件における学習過程と得点推移

4 まとめと今後の展望

本研究では報酬値に含まれるノイズを考慮した学習安定化を目的に，状態価値の分散を予測に関する特徴マップを利用した深層強化手法を提案した．Open AI Gym の Atari において評価実験を行い，提案手法の有効性が確認された．今後の展望として，より多様なタスクによる検証やロボットタスクに応用した検証を行う．

謝辞

本研究は，JST，ACT-X，JPMJAX190I の支援を受けたものである．ここに謝意を表します．

参考文献

[1] S.Levine, et al., ”Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” Int. J. Robot. Res., vol.37, no.4–5, pp.421–436, 2018.
[2] D.Silver, et al., ”Mastering the game of Go with deep neural networks and tree search,” Nature 529, 484–489, 2016.
[3] E.Jang, et al., ”Grasp2Vec: Learning Object Representations from Self-Supervised Grasping,” in Conf. on Robot Learning (CoRL), 2018.
[4] K.Suzuki, et al., ”Online Self-Supervised Learning for Object Picking: Detecting Optimum Grasping Position Using a Metric Learning Approach,” in Proc. of Int. Symp. on System Integrations, 2020.
[5] B.O.Donoghue, et al., ”Reinforcement Learning with a Corrupted Reward Channel,” in 26th International Joint Conf. on Artificial Intelligence (IJCAI-17), 4705–4713, 2017.
[6] J.Romoff, et al., ”Reward Estimation for Variance Reduction in Deep Reinforcement Learning,” in arXiv preprint arXiv:1805.03359, 2018.
[7] V.Mnih et al., ”Asynchronous methods for deep reinforcement learning,” in Int. Conf. on Machine Learning (ICML), 2016.
[8] S.Murata, et al., ”Learning to reproduce fluctuating time series by inferring their time-dependent stochastic properties: Application in robot learning via tutoring,” IEEE Transactions on Autonomous Mental Development, 5, 298310, 2013.
[9] H.Fukui, et al., ”Attention Branch Network: Learning of Attention Mechanism for Visual Explanation,” in Conf. on Computer Vision and Pattern Recognition (CVPR), 2019.
[10] 福井, 他, ”Attention 機構を導入した A3C の提案,” 日本ロボット学会学術講演会, 2018.