ISSN 2186-5647

P-17

内部報酬を付加した探索強化型 A2C

日大生産工　○松原　忠司　　日大生産工　山内　ゆかり

1. まえがき

強化学習にはQ-LearningやSARASAなど様々な手法が存在する。その一つであるActor-Criticの応用としてAdvantage Actor Critic：A2C+Contingency-aware Exploration：CoEX[1]が2019年に提案された。A2Cとは、Advantage関数を用いてActor-Criticを行うことで学習の安定を図る学習法である。そこにAttentive Dynamics Model：ADMでエージェントが取る行動を予測することで、A2Cに偶発性を取り入れる。それにより、Atariゲームにおいて11,000以上という非常に高いスコアを獲得に成功している。だが報酬が複数あるときに局所解に陥ってしまう問題がある。

本研究では、内部報酬を考慮ことで未探索の環境であっても積極的に行動する手法を提案し、強化学習における計算機実験により提案手法とA2C+ CoEXを比較し、局所解に陥る問題を解決しているのかについて報告する。

2. 従来研究

2-1 Advantage Actor Critic

Advantage Actor Critic：A2CはAsynchronous Advantage Actor Critic：A3C[2]の派生手法である。A3Cでは複数のエージェントを同じ環境下で非同期に学習し、求めた勾配情報から分散学習を行う学習法である。この手法から非同期の機能を削ることでGPUの数とエージェントの数が同一である問題を解消することが出来る。つまり複数のエージェントから得た結果を一つのエージェントにまとめて学習することで更新する学習法になる。A2Cの概要を図１に示す。

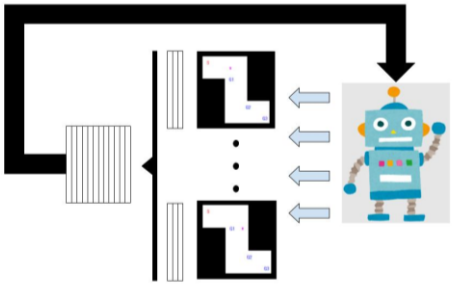

図１　A2C の概要

この手法ならGPUの数を１つで抑えることができるため、リソースを大きく抑えることができる。

A2CではNeural Networkの機能を搭載したActor-Criticとadvantage関数を使用する。advantage関数はある状態から特定の行動を選択したときの価値から、ある方策に従って状態を選択した際の価値を引いた値である。現在の状態をs_t、行動をa_t、γを割引率、報酬をR_{t+1}、現在の状態価値関数を$V(s_t, a_t)$としたときのadvantage関数$A(s_t, a_t)$を式(1)に示す。

$$A(s_t, a_t) = \left(\sum_{i=0}^{N-1}\left(\gamma^i R_{t+i+1}\right)\right) + \gamma^N V(s_{t+k}) \quad (1)$$

このadvantage関数と損失関数Lを用いた計算をすることで、更新に用いられる推定量$G(\omega)$を作ることが出来る。学習率をαとしたとき$G(\omega)$を式(2)に示す。

$$G(\omega) = A(s_t, a_t) + \alpha L \quad (2)$$

(2)で更新を行うことで強化学習の効率を上昇させる。A2CはA3Cに劣らない性能を発揮できる上に、A3Cより少ないリソースで実行することが出来る。

2-2 Attentive Dynamics Model

Attentive Dynamics Model：ADMは偶発性を活用し、環境における状態から選択されるエージェントの行動を予測する。2つの連続する入力フレームs_{t-1}、$s_t \in S$を入力とし、エージェントがs_{t-1}からs_tへ行動を使用して予測する。概要を図2に示す。

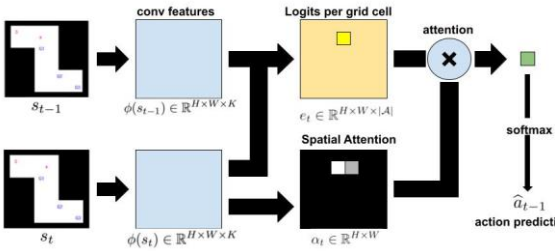

図２　ADM の概要

Exploration-Enhanced A2C with Internal Rewards

Tadashi MATSUBARA and Yukari YAMAUCHI

図２の手順をたどることで学習における偶発性を発生させる。そして出力した偶発性$\hat{a}_{t-1}$を学習に取り入れることで学習に影響を与える。

2-3 Random Network Distillation

Random Network Distillation：RND[3]はPredictor NetworkとTarget Networkという２つのネットワークを用いる。Target Networkは学習を行わず、Predictor NetworkはTarget Networkの出力に近づくように学習をする。これらのネットワークの予測誤差を活用することで内部報酬R^Iを生成する。R^Iを式(3)に示す。

$$R^I = |\hat{f}(x) - f(x)|_2^2 \quad (3)$$

RNDの内部報酬生成の様子を図3に示す。

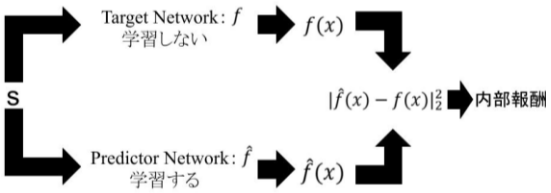

図３　RND の概要

学習が早期のうちは必然的に予測誤差の値が大きくなるため、内部報酬が大きくなる。そのため未探索の場所も積極的に探索する。学習が進み予測誤差の値が０に近づくと内部報酬も０に近づくため、外部環境が獲得可能となる。そのため外部報酬を長期的に最大化する方策を獲得することが可能となる。

3.　提案手法

ADMは偶発性に頼っているため、複数の報酬があったときに一番大きい報酬を受け取れない問題がある。そこで内部報酬を活用した強化学習を活用することで、未探索の領域にも積極的に探索を促す。今回はRNDを用いて内部報酬を探索する。提案手法として、A2C+CoEXで計算するアドバンテージ関数に内部報酬を加味することで、局所解に陥る問題を解決する。内部報酬R_t^Iから得たアドバンテージ関数$A^I(s_t, a_t)$を式(4)に示す。

$$A^I(s_t, a_t) = \sum_{i=0}^{N-1} \left(\gamma^i R_{t+i+1}^I\right) + \gamma^N V^I(s_{t+k}) \quad (4)$$

この$A^I(s_t, a_t)$と外部報酬から得たアドバンテージ関数$A^E(s_t, a_t)$を足した$A(s_t, a_t)$を式(5)に示す。

$$A(s_t, a_t) = A^I(s_t, a_t) + A^E(s_t, a_t) \quad (5)$$

この$A(s_t, a_t)$最終的なアドバンテージ関数として計算する。既存のアドバンテージ関数に内部報酬を加味することで局所解に陥る危険性を大きく減らすことができる。

4.　実験

本研究は12×12の複数の報酬が存在する迷路問題を使用する。実験環境を図4に示す。

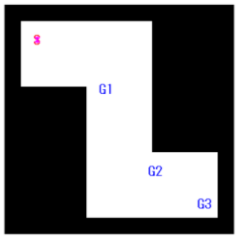

図４　実験環境

G1、G2の報酬を0.1としG3の報酬を1.0とする。図4の環境で1000回学習を行うことで、最も報酬の値が大きいG3にエージェントが収束しているか確認する。

5.　まとめ

本研究では従来研究の問題点であった複数報酬での局所解を、好奇心を導入することで解決を図った。結果として、従来研究よりも累積報酬が増加し、学習結果も同様の結果を得ることが出来た。しかし、本研究の従来研究は偶発性の制御をしていない。また環境が複雑なものではないため、好奇心の必要な場面を作りづらい問題点がある。そのため今後偶発性を制御することで提案手法と差が出来るかについて確認すると共に、よりまばらな報酬で学習できる環境で実験を行うことで結果を確認する。

参考文献

[1] J.Choi, Y.Guo, M.Moczulski, J.Oh, N.Wu, M.Norouzi and H.Lee “Contingency-Aware Exploration in Reinforcement Learning”, arXiv(2019)

[2] V.Mnih, A.P.Badia, M.Mirza, A.Graves, T.Harley, T.P.Lillicrap, D.Silver and K.Kavukcuoglu “Asynchronous Methods for Deep Reinforcement Learning”, Proceedings of Machine Learning Research (2016)

[3] 岩科 亨, 森山 甲一, 松井 藤五郎, 武藤 敦子, 犬塚 信博, マルチエージェント強化学習問題への好奇心探索の適用, 人工知能学会全国大会論文集,(2021), p1
Online ISSN：2758-7347