ISSN 2186-5647

P-10

好奇心探索を用いたDDPG

日大生産工　○田中　隼也　　日大生産工　山内　ゆかり

1. まえがき

従来、決定論的方策勾配法(Deterministic Policy Gradient Algorithms：DPG)は存在しないと考えられていた。Silverら[1]は、DPGが実際に存在することを提案し、確率論的方策勾配法(Stochastic Policy Gradient Algorithms：SPG)よりも効率的に計算できることを検証した。そして、DPGを改良したものが深層型決定論的方策勾配法(Deep Deterministic Policy Gradient Algorithms：DDPG)[2]である。DDPGは収束が早く、複雑な環境下で学習できるが、局所解に陥りやすい。

そこで、Burdaら[3]が提案した好奇心探索を用いることで、DDPGにおける問題点の解決を試みる。好奇心探索とは、エージェントの過去の経験に基づいて学習されたネットワークの誤差である内部報酬を使用して、新しい経験の新規性を定量化することができる手法である。本研究では、局所解に陥りやすいというDDPGのデメリットに着目し、解決する提案として好奇心探索を導入する。

2. 従来研究

2.1 Actor Critic

Actor Criticとは、行動を選択するActorと、Actorが選択した行動を評価するCriticで構成される強化学習の一つである。

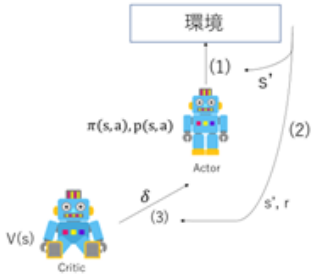

図１　Actor Critic の概要

Actor Criticの学習の流れは図1に示す。図1の(1)では、Actorは方策π(s, a)を元に行動を選択し、行動をする。(2)では、環境から状態s'及び報酬rがCriticに渡される。(3)では、得られた状態s'、報酬rを使ってActorの取った行動の評価を行い、Actorに渡す。Actorは評価を元に方策の更新をする。これらを繰り返し行う。

2.1.1 Actor

Actorは実際の行動を決定して実行する。実際に行動した結果に関しては、Criticによって評価される。Criticによって評価された結果を使って自らの行動価値を修正する役割を持つ。

2.1.2 Critic

Criticとは、評価器と呼ばれているが評論家という意味で、Actorを批評する役割を持つ。そして、得られた報酬や遷移先の状態を用いてTD誤差δを計算する。

2.2 確率論的方策勾配法(SPG)

方策勾配法とは、連続行動空間での強化学習によく用いられる手法の一つであり、確率的方策をモデル化している。この方策勾配法は、以下の方策勾配定理に基づいて計算される。

$$\nabla_\theta J(\pi_\theta) = \int_S \rho^\pi(s) \int_A \nabla_\theta \pi_\theta(a|s) Q^\pi(s,a)\, dads \quad (1)$$
$$= E_{s\sim\rho^\pi,\ a\sim\pi_\theta}[\nabla_\theta log\pi_\theta(a|s) Q^\pi(s,a)]$$

2.3 決定論的方策勾配法(DPG)

決定論的方策勾配法とは、方策勾配法の行動方策を更新する手法であり、価値関数の勾配を取ってその方向に方策を変化させている。このDPGは、以下の決定論的方策勾配定理に基づいて計算される。

$$\nabla_\theta J(\mu_\theta) = \int_S \rho^\mu(s) \nabla_\theta \mu_\theta(s) \nabla_\alpha Q^\mu(s,a)|_{\alpha=\mu_\theta(s)} ds \quad (2)$$
$$= E_{s\sim\rho^\mu}[\nabla_\theta \mu_\theta(s) \nabla_\alpha Q^\mu(s,a)|_{\alpha=\mu_\theta(s)}]$$

式(1)、式(2)から、DPGでは方策に関して期待値計算が必要ないため、SPGよりも効率的に計算が可能である。

2.4 Deep Q-Network(DQN)

DQNとは、Q学習に深層学習(Deep Neural Network：DNN)の考え方を含めた手法である。DNNとは、隠れ層が複数存在するNNによって重みを更新していく手法である。

2.5 深層型決定論的方策勾配法(DDPG)

DDPGとは、DPGをベースにDQN(Deep Q-Network)を組み合わせたものである。問題が複雑になるとDPGでは結果が不安定になるので、

DDPG with Curiosity Search

Shunya TANAKA and Yukari YAMAUCHI

DDPGが提案された。式(3)はDDPGの更新式である。

$$\nabla_{\theta^\mu} J = \frac{1}{N}\sum_i (\nabla_a Q(s,a|\theta^Q)|_{s=s_i,\ a=\mu(s_i)} \nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s_i}) \quad (3)$$

2.6 好奇心探索

好奇心探索[4]とは、未知の状態へ行動することで得られる、環境から得られるものとは別の報酬である内部報酬を設定することで、未知の状態を積極的に探索する方策を得る手法のことである。ただし、環境から得られる報酬を最大化する方策を得ることが目的なので、内部報酬は最終的に発生しないようにする。

好奇心探索の手法の一つとしてRandom Network Distillation(RND)が存在する。RNDは状態を入力とする2つのネットワーク「Predictor Network」と「Target Network」を用いる。Target Networkは学習を行わないが、Predictor NetworkはTarget Networkの出力に近づくように学習する。この2つの予測誤差を利用して内部報酬を生成する。

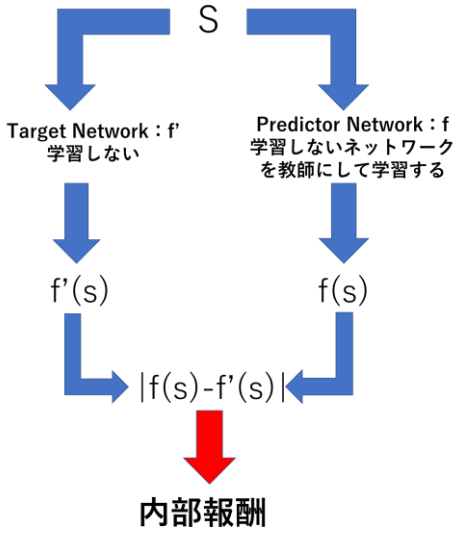

図2　内部報酬の生成

内部報酬の計算方法については図2に示す。状態sの経験が少なければ学習が進んでいないので内部報酬が大きくなり、積極的に探索するようになる。状態sの経験が増加すると、学習が進み内部報酬が減少する。最終的に、目新しい状態がなくなり、内部報酬が0に近づく。よって、本来の環境から得られる外部報酬を長期的に最大化する方策を獲得可能である。内部報酬を取り入れた更新式は式(4)となる。ここで内部報酬はi_tである。

$$Q(s_t,a_t) = Q(s_t,a_t) + \alpha(r_t + i_t + \gamma \times maxQ(s_{t+1},a) - Q(s_t,a_t)) \quad (4)$$

3. 提案手法

本研究では、DDPGは局所解に陥りやすいというデメリットに焦点を置く。そこで、好奇心探索を導入することで、内部報酬を用いて未知の状態を探索していく手法を組み込み局所解に陥ることを防ぐ。

4. 実験および検討

DDPGに好奇心探索を導入することで、未知の状態においての探索を積極的に行うことになる。よって、経路が全体の経路を参照して確定するので、DDPGのデメリットである局所解に陥るという行動が減少するのではないかと考えられる。

5. まとめ

本研究では、DDPGに好奇心探索を導入することを提案した。提案手法では、未知の状態を積極的に探索する方策を得る手法を用いている。このことから、DDPGのデメリットである局所解に陥りやすいという問題点を解決することができると考えられる。

参考文献

[1]David Silver、Guy Lever、Nicolas Heess、Thomas Degris、Daan Wierstra、Martin Riedmiller、“Deterministic Policy Gradient Algorithms” ICML'14:Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 31・June 2014・Pages I-387–I-395

[2]Timothy P、Lillicrap、Jonathan J. Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver & Daan Wierstra、“CONTINUOUS CONTROL WITH DEEP REINFORCEMENT LEARNING” Published as a conference paper at ICMR 2016

[3]Yuri Burda、Harrison Edwards、Amos Storky、Oleg Klimov、“EXPLORATION BY RANDOM NETWORK DISTILLATION” arXiv:1810.12894v1 [cs.LG] 30 Oct 2018

[4]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応” The 35th Annual Conference of the Japanese Society for Artificial Intelligence, 2021

[5]曾我部東馬、”強化学習アルゴリズム入門「平均」からはじめる基礎と応用”（2019）