ISSN 2186-5647

A3C における知覚情報の粗視化の導入

日大生産工　○植竹　青海　　日大生産工　山内　ゆかり

1. まえがき

強化学習はシステム自身が試行錯誤を繰り返し、最適解を導き出す学習手法である。この強化学習に対して飛躍的技術進展をもたらしたのが、強化学習と相性のいい深層学習とを組み合わせた深層強化学習である。深層強化学習で有名なものとしてDeep Q-network (DQN)や2016年に登場したAlphaGo(アルファ碁)があげられる。だがこれらの強化学習手法は、学習に多くのデータや学習時間を要する問題がある。

MnihらはAtari2600のゲーム画面を入力、ゲームスコアを報酬として、深層強化学習を行うことで熟練プレイヤーのスコアを凌駕することに成功した。その際、Asynchronous Advantage actor-critic[1] (A3C)が提案された。この手法は複数のエージェントを用いて、それぞれが探索し、行動選択の方策を評価して更新する。複数エージェントを用いることで時間の大幅な削減ができている。しかしA3Cでは複数のエージェントがいることで知覚する状態数が多くなり、メモリと計算量が多く必要となるという問題がある。

本研究では 知覚情報の粗視化[2]の導入を提案し、追跡問題における計算機実験により提案手法とA3C、Actor-Critic[3]の3つを比較し、計算量の削減について報告する。

2. 従来研究

2.1　強化学習

強化学習とはSuttonらが研究し提案した学習手法である。エージェントが環境から状態を認識し行動選択を行い、ゴールした時に最大の報酬を得られるようにする学習方法である。

2.2　Actor-Critic

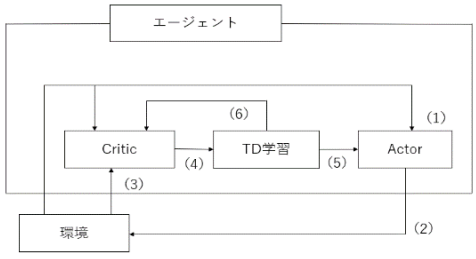

Figure 1　Actor-Critic のモデル

Actorは環境から状態を認識し、与えられている確率分布に基づき行動を選択する。CriticはActorの行った行動に対して評価を行い、状態価値関数の更新を行う。Actor-Criticでの更新の有無はTemporal Difference Learning[4](TD学習)を利用する。このTD学習ではActorとCriticの両方を修正し、1stepの行動でエージェントの状態変化に伴う状態価値関数を変化させる。

(1),(2)はTD学習による更新式で行動価値を更新する。方策の更新式は(3),(4)である。(3)は行動した選択が良い結果だった場合の更新式である。(4)はそれ以外の選択されなかった行動の更新式である。i=1,2,3である。

$$r_{t+1} \quad +\gamma V(S_{t+1}) - \ V(S_t) \tag{1}$$

$$V(S_t) + \alpha(r_{t+1} \quad +\gamma V(S_{t+1}) - V(S_t)) \tag{2}$$

$$a_0 = \frac{C + a_0}{C + a_0 + a_1 + a_2 + a_3} \tag{3}$$

$$a_i = \frac{a_i}{C + a_0 + a_1 + a_2 + a_3} \tag{4}$$

2.3　Advantage

Advantageは1step先だけでなく、複数step先まで動かして更新するという考え方である。A3CはAsynchronousとAdvantageの要素、そしてActor-Criticを取り入れた学習法である。

$$A^{\pi}(s_t, a_t) = \sum_{n=0}^{N-1} r_{t+n} + V^{\pi}(S_{t+N}, a_{t+N}) - V^{\pi}(s_t, a_t) \tag{5}$$

(5)はN-step学習の式である。

2.4　A3C

A3Cは複数のエージェントを並列で走らせActor-Critic法（TD学習）を使用し、その結果を用いて行動価値を更新し、最適な行動をさせるという学習手法である。最大の行動回数まで達するか、また報酬を獲得するまでエージェントを行動させる。その後得られた結果をもとに、各行動履歴を遡り方策と、行動価値を更新する。

2.5　状態数削減

状態数はエージェントの状態行動に比例するためエージェントが多いほど状態数は多くなる。そのため学習速度の低下やメモリへの負担、学習時間に大きく関わる。この状態数を抑えるには知覚情報の粗視化という手法が用いられる。

これは知覚情報を粗くすることで探索すべき状態数を削減する方法である。追跡問題で扱う場合には位置情報を上下左右とその中間を合わ

Introduction of Rough Representation of Perceptual Information in A3C
Haruka UETAKE and Yukari YAMAUCHI

せた領域で分類し状態数を削減する。つまり今回は位置情報を制限した削減方法で行う。

3. 提案手法

本研究では A3C における計算量の減少を目的として、知覚情報の粗視化の導入を提案する。提案手法ではエージェント数が多くなると発生する状態数の増加、それに伴う計算量の削減になる。Figure2 の緑の点がエージェントで、数字の書いてある順にその範囲を、エージェントや敵がいるのかを判断する。

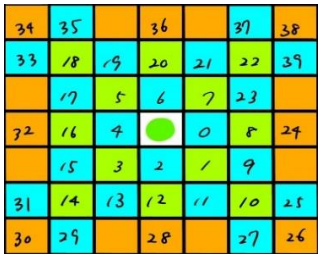

Figure 2　知覚情報の粗視化

4. 実験および検討

今回の実験の実験環境は 7×7 の Grid World の追跡問題である。エージェントを複数用意して 1step ごとに上下左右のいずれかの方向に行動、また「何もしない」の 5 パターンである。この実験では Figure3 の赤い点を敵としており緑色のエージェントが 3 方向から囲むことができればクリアとなる。今回提案した提案手法と A3C、Actor-Critic の 3 つを比較し計算量の削減について求める。

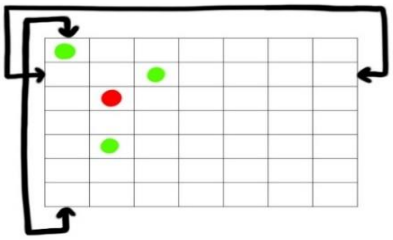

Figure 3　追跡問題のモデル

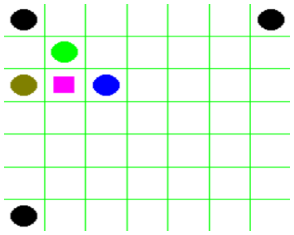

Figure 4　実験環境の動き

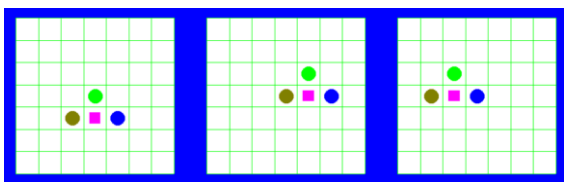

Figure 5　各エージェントの相対位置

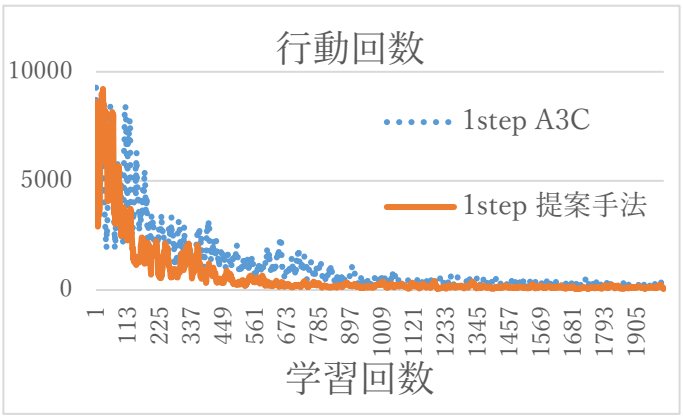

Figure 6　学習曲線

Figure6 は 1step で行った A3C と提案手法の学習回数 2000 回、行動回数 50000 回の学習曲線である。

Table 1　各手法の行動回数の合計

	1step	3step
A3C	2344727回	6665677回
提案手法	1339394回	5799766回

5. まとめ

本研究ではエージェントが多くなることでかかる計算量や計算時間の減少を目的として、状態数を減らすために知覚情報の粗視化の導入を提案した。Table1の行動回数の合計を比較すると、A3Cのみよりも提案手法を導入した結果の方が、大幅に減少している。だが、A3Cの特徴であるAdvantageを使用した3stepの結果は1stepよりも行動回数が多く、提案手法とは相性が悪いのではないかと考えられる。

参考文献

1) Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves ,Tim Harley, Timothy P. Lillicrap, David Silver and Koray Kavukvuoglu, “Asynchronous Methods for Deep Reinforcement Learning” Proceeding of the 33rd International Conference on Machine Learning. (2016) vol.48

2)伊藤僚,吉川毅,野中秀俊,マルチエージェント強化学習における知覚情報の適応的粗視化,26th Fuzzy System Symposium September 13-15,2010

3) 三上貞芳,皆川雅章,強化学習,森北出版株式会社,pp.161-164

4) 三上貞芳,皆川雅章,強化学習,森北出版株式会社,p.142