ISS-P-023

自動運転モデルの構築を目指したActor-Critic法の検討

中川　将輝† 八上　剛† 渡辺　秀行‡ 片桐　滋† 大崎　美穂†
†同志社大学 ‡ATR

1. はじめに

自動車の自動運転技術開発の一環として，遺伝的プログラミングを用いた自動運転モデルの構築が行われている[1]．そこでは，優れたモデルが得られる一方で，学習に膨大な時間を要する問題が示されている．

本研究では，強化学習の一つであり，状態空間及び行動空間が連続量であるタスクに対しても良い方策を獲得できると報告されている Deterministic Policy Gradient（DPG）による Actor-Critic 法[2]を用いて，先行研究[1]と同様に車線復帰を行う自動運転モデルの構築を行い，その性能を実験的に評価する．

2. 強化学習

2.1 Actor-Critic 法

Actor-Critic 法は，観測された状態から方策に基づき行動を決定する Actor と，Actor が選択した行動を価値関数を用いて評価する Critic の2つで構成されている．そこでの学習の目標は，式(1)に示す各時刻で得られる報酬の累積R_tが最大となるような方策を Agent が獲得することである．

$$R_t = \sum_{i=t}^{\infty} \gamma^{i-t} r_t \,, \tag{1}$$

ここで，γは割引率，r_tは時刻tにおいて得られた報酬を表す．一般的な Actor-Critic 法は，Temporal difference error（TD 誤差）を用いて Actor の方策関数更新と Critic の価値関数更新を行うが，価値関数はルックアップテーブルで表現されているため，少なくとも状態空間が連続量である場合，離散化を行う必要がある．そこで本研究では，状態空間を離散化することなく効率的に学習が行える DPG による Actor-Critic 法を採用する．

2.2 DPG を用いた Actor-Critic 法

DPG を用いた Actor-Critic 法は，Actor の方策関数と Critic の価値関数を共にニューラルネットワーク（NN）で表現する．Actor 側と Critic 側の NN のパラメータをそれぞれθ^{μ}とθ^{Q}とおくと，Critic のパラメータθ^{Q}は式(2)の損失関数Lを最小化するように更新される．

$$L = \frac{1}{N}\sum_{i}\left(y_i - Q(s_i, a_i; \theta^{Q})\right), \tag{2}$$

$$y_i = r_i + \gamma Q\left(s_{t+1}, \mu(s_{t+1}; \theta^{\mu}); \theta^{Q}\right), \tag{3}$$

また，Actor のパラメータθ^{μ}は Policy Gradient と呼ばれる式(4)を用いて更新される[2]．

$$\nabla_{\theta^{\mu}} J \approx \frac{1}{N}\sum_{i} \nabla_a Q(s, a; \theta^{Q})|_{s=s_i, a=\mu(s_i;\theta^{\mu})} \nabla_{\theta^{\mu}} \mu(s; \theta^{\mu})|_{s=s_i} \,, \tag{4}$$

ここで，s_tとa_tはそれぞれ時刻tにおける状態と行動を，$Q(\cdot)$は行動価値関数，$\mu(\cdot)$は方策関数，Nはパラメータ更新時に用いるバッチサイズを表す．

3. 評価実験

3.1 概要

実験環境として，運転シミュレーターTORCSを用いた．直線道路の中央線に沿って走行している自動車に対し，故意にステアリングを右に一定時間切り，その後一定時間を Agent がステアリングを操作して車線復帰するという課題設定でシミュレーションを行った．報酬は式(5)のように定義した．

$$r = -\left(\frac{d}{d_{MAX}}\right)^2 - \left(\frac{a_y}{a_{yMAX}}\right)^2 , \tag{5}$$

ここで，dは中央線と自動車との距離，a_yは自動車にかかる横向き加速度，d_{MAX}とa_{yMAX}は報酬の値を正規化するために設けた定数である．

3.2 実験結果

上記の車線復帰の（学習における）試走を 50 回繰り返して得られた Agent の動作結果を図 1 に示す．

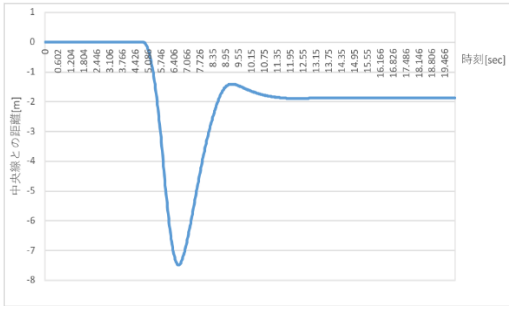

図 1 学習後における Agent の動作結果

4. おわりに

比較的短時間の学習で車線復帰に向けた運転モデルを獲得できたが，学習の安定性とその質にはまだ改良が必要であった．引き続き，改良を試みる予定である．

謝辞：本研究の一部は文科省 H26 年度私学戦略的研究基盤形成支援事業「ドライバ・イン・ザ・ループ」の支援を受けて行われた．

参考文献

[1] Go Yakami, et al.: “Automobile Driving Support System Evolved by Genetic Programming”, Proc. TENCON, Nov. 2016

[2] Lillicrap, P.T., et al.: “Continuous Control with Deep Reinforcement Learning”, Proc. ICLR,2016, 2016.