3I2-OS-13b-5

Deep Q-Network による RC カー群の運動制御を実現する協調学習の提案

Proposal of Cooperative Learning to Realize Motion Control of RC Cars Group by Deep Q-Network

小川一太郎[*1] 横山想一郎[*1] 山下倫央[*1] 川村秀憲[*1] 酒徳哲[*2] 柳原正[*2] 田中英明[*2]
Ichitaro Ogawa Soichiro Yokoyama Tomohisa Yamashita Hidenori Kawamura Akira Sakatoku Tadashi Yanagihara Hideaki Tanaka

[*1] 北海道大学情報科学院
Graduate School of information Science and Technology, Hokkaido University. #1

[*2] KDDI 総合研究所
KDDI Research, Inc. #2

We propose a learning method using Deep Q-Network aiming at acquiring autonomous driving by multiple RC cars. In addition to not colliding with other vehicles and not going out of the track, the RC cars aim to turn to the direction specified at the intersection. LIDAR information is converted from the position and direction of the vehicle obtained from the image of the camera attached to the ceiling, and learning is performed based on the data with 1 episode running from the start to the end of traveling. As a result of the experiment learning to increase the mileage was done but no learning to increase the route selection accuracy rate was done. Investigation of reward design that increases both indices is necessary.

1. 研究背景

近年, 様々なロボットが機械学習を用いて動作を学習している. 実環境中のロボットが複雑な環境を認識し, 動作を決定することをニューラルネットで行う研究がなされている.

Deep Q-Network[Mnih 2015]は状態を評価して行動を決定するという面で, エージェントに行動をさせて学習させることに適している. 例えば, ビデオゲームを学習させて人間の操作よりも大きくスコアを獲得した例や, 自動車の交通の整理[Li 2016]など様々な研究がされている.

一方で, Deep Q-Network を用いた際にロボットの行動価値を評価する報酬の設計は, 期待している動作をさせるためにどんな行動に対してどれだけの報酬を与えればよいのか判断することが難しい. また, 複数のエージェントに適応する場合には特に問題となり, エージェント同士での行動を認識し, 報酬の設計を正確に行う必要がある. 例えば自動車などでは車車間通信にて入手した情報をどのように使用するか検証する必要がある.

複数のエージェントが行動する例として, 本研究では RC カーの走行を取り上げる. RC カーの動作に必要な入力はアクセルとステアリングの二つとする. また, RC カーが走行するために環境を認識する必要がある. モデル化が行い易く, ニューラルネットのサイズが大きくなりすぎないようにするために, コースの情報だけを環境の入力とする設定にした.

RC カーの運転制御学習を行うため, 走行に必要な要素に対して Deep Q-Network の報酬を設定し, 複数の RC カーの制御を目指した.

Preferred Networks[PFN 2016]も RC カーを用いた自動運転の研究を行っており, こちらでは走行コースに沿った走行を実現している. 本研究では外部からの命令によって交差点内の進路を決定し走行する.

2. 実験システム

RC カーの学習環境を作成する. 本研究の学習では, コースアウトや他 RC カーとの衝突を起こさず交差点での進路を入力によって制御できることを目標としたため, これが実現可能な環境を作成する. 学習環境を説明するにあたって, 大きく分けて 3 つの構成に分けることができる.

連絡先: 小川一太郎, 北海道大学情報科学研究科, ogawa_ogawa@complex.ist.hokudai.ac.jp

図 1 RC カーとマーカー

2.1 走行部

RC カーを制御するための機構. RC カーにはマイコン(ラズベリーパイ 3)を搭載し, 制御を行う. RC カーの受信機ではなくマイコンからパルス信号を送るようにすることでシステムによる制御を可能とした.

走行用のコースは 5.46*m*×7.28*m* の石膏ボードを用意し, システムの処理では 1092*px*×1456*px* となるようにする. 対角線上に二台のカメラを設置し, RC カー上部のマーカーを画像から位置や方向を認識する. 走行可能な領域の幅は 1*m* とした.

2.2 走行制御部

RC カーの位置を推定するため, 固定されたカラーカメラよりマーカーを認識する. 処理速度を保つためカメラは二台, コースの対角に俯瞰するように設置した. マーカーは RC カー上に緑もしくは黄の円形シールを 3 つ貼り, 色毎にカメラ画像を閾値処理した結果位置と方向が求められる. 検知には粒子フィルタ[北川 1996]を導入した. システムのループは 0.15 秒毎に行っている.

運転制御の学習のため入力出力を決定した.

まず入力は RC カー本体の情報である, 速度[m/s], 加速度[m/s^2], 角速度[rad/s], 角加速度[rad/s^2], 直前 3 ステップのハンドル角とアクセル. 交差点での進路を設定するため, 進むべき道を目標方向として RC カーの進行方向との誤差を入力に追加する. コースや他 RC カーを検知する方法としてリモートセンシング技術であるライダーを仮想的に導入した. コース検知用のライダーと RC カーの検知用のライダーを設定した. RC カーの前方 90 度に 32 方向, 他 270 度に 24 方向にそれぞれライダーを発する. ニューラルネットの入力の合計数は 123 である.

次に出力では RC カーがとることができる行動を設定する. 直進と左右それぞれ三種類(ハンドル角約 7 度, 22 度,30 度)を設定し, コーナーの最小半径である 0.405*m* を旋回でき, 微調整ができるようにした. さらに交差点での衝突を回避するために停止を設定した. 出力の合計数は 8 である.

2.3 学習部

学習部にはデータベースの一つである MySQL で走行データが保存されている.

走行データ一つにつき, 走行の開始からコースアウト, 他 RC カーとの衝突または 100 ループ中に 80 回停止したときまでの入力出力データを 1 エピソードとして保存する.

tensorflow を用いて Deep Q-Network を実装した.

3. 学習実験

今回運転制御学習のため設定した報酬は次の通り.

(1) コースアウトで減点

コースアウト時に-2900 の報酬, エピソードを終了する.

(2) 他 RC カーとの衝突で減点

衝突時に-10000 の報酬, エピソードを終了する.

(3) 停止に対して減点

行動のうち, 停止を採択したときに-50 の報酬.

(4) ハンドルの変更に対して減点

7 段階あるハンドル角が変更された際, 変化の大きさに応じて-0.5, -2.5, -5, -35, -45, -50 の報酬.

(5) 目標方向との誤差に加点減点

目標方向と RC カーの進行方向との差の 1 ステップの変化量に対して, 標準化した後-1200/π を掛けた報酬.

(6) ライダーの長さに加点減点

左車線の走行をするために設定する. RC カーの左 0.25*m* 右 0.75*m* 前方 1*m* を通るように左右それぞれに楕円を描き, コース検知用のライダーをその楕円と比較して与える. 標準化した値に 70 を掛けた報酬.

4. 学習実験

実際に学習を行い, その動作を調査した. カリキュラム学習[Bengio 2009]を導入し, 学習に段階をつけた. 今回では単機走行, コース上に停止車を配置した走行, 二機走行とした.

単機走行では図 2 のコースを使用した. 停止車を配置した走行では図 2 のコースの直線部分に 4~5 台仮想的に配置した. 二機走行の前半 50 万学習は交差点でのすれ違いをより学習し易くするため図 3 の八の字コースで走行を行った.

結果はエピソードごとの走行距離(図 4)と交差点でのルート選択正解率(図 5)の 5 万学習中の平均をプロットした. 学習が 300 万回までが単機での走行, 400 万回までが停止車を配置した走行, 500 万回までが二機での走行である.

走行距離のグラフがグラフ 1, 交差点でのルート選択正解率のグラフがグラフ 2 である. 走行距離を大きく上昇させる学習が行われているがルート選択正解率を上昇させる学習は行われていないことが分かる.

5. まとめ

今回の実験では走行距離を伸ばす学習が行われ, ルート選択の正解率は上昇しなかった. ルート選択正解率が伸び悩んでいることから目標方向の報酬設定が不適切であったと考えられる. 走行距離とルート選択正解率を同時に上昇させるように学習を行うには目標方向と RC カーの方向との差に対する報酬の重みを変化させて適切な値を調査するほか, ルート選択が正解した時に加点を行うような直接的な報酬設定などに変更するなど, 報酬設定を調査する必要がある.

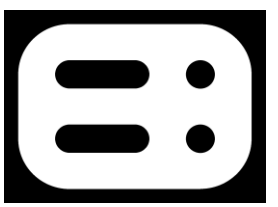

図 2 コース 1

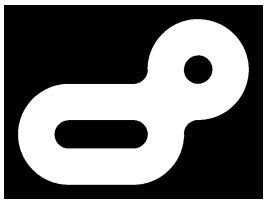

図 3 コース 2

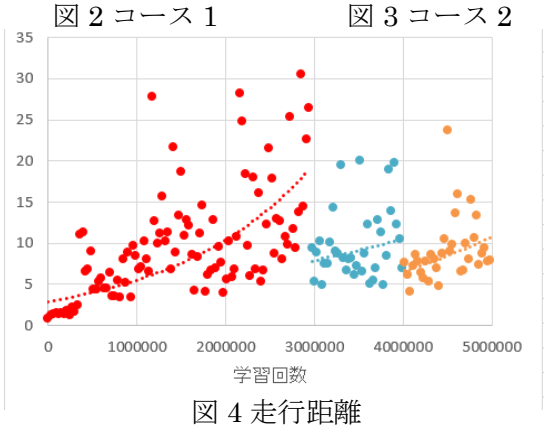

図 4 走行距離

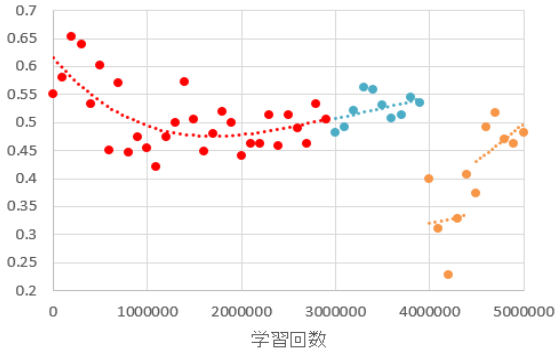

図 5 ルート選択正解率

参考文献

[Mnih 2015] Volodymyr Mnih.. , Koray Kavukcoglu. , David Silver, . : Human-level control through deep reinforcement learning, Nature, Nature Publishing Group ,2015.

[Li 2016] Li Li. , Yisheng Lv. , Fei-Yue Wang. : Traffic Signal Timing via Deep Reinforcement Learning , IEEE/CAA JOURNAL OF AUTOMATICA SINICA, IEEE , 2016.

[PFN 2016] Preferred Networks : CES2016 でロボットカーのデモを展示してきました , Preferred Reserch, https://research.preferred.jp/2016/01/ces2016/, 参照日 2016/2/28(2016)

[北川 1996] 北川源四郎. : モンテカルロ･フィルタ及び平滑化について, 統計数理, 統計数理研究所, 1996.

[Bengio 2009] Yoshua Bengio. , Jérôme Louradour. , Ronan Collobert . , Jason Weston . : Curriculum Learning , ICML'09 Proceedings of the 26th Annual International Conference on Machine Learning, 2009.