100 ms周期で走行車線を選択可能な深層強化学習手法

Deep Reinforcement Learning Method Allowing Selection of Automobile Driving Lane in 100-Millisecond Cycle

短い行動決定周期に対応した深層強化学習手法で，一般道での自律走行の実現を目指す

近年，一般道のような複雑な環境下で自律走行を実現するため，従来のルールベースの手法による自律走行計画技術に替わって，深層強化学習手法を用いた自律走行計画技術が検討され始めています。しかし，従来の深層強化学習手法は，行動の決定を行う周期が短いと十分な学習ができず，俊敏かつ安全な自律走行を実現できないという問題がありました。

東芝は，このような問題を解決するために，行動選択の一貫性を考慮して行動探索を行う独自の強化学習手法を開発し，自動車向けの画像認識と同じ実行周期100 msでの走行車線の選択を学習可能にしました。これによって，一般道での自律走行の実現に向けた更なる技術進展が期待されます。

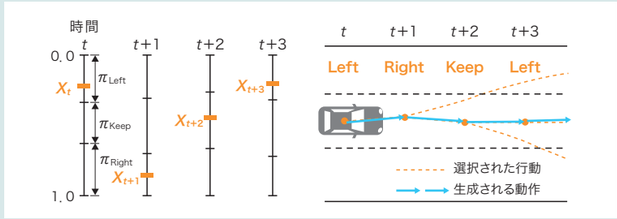

図1．従来の行動探索手法
各時間ステップでの行動は，一定周期ごとに独立に選択されるため，結果的に車線維持に近い動作ばかりが発生します。

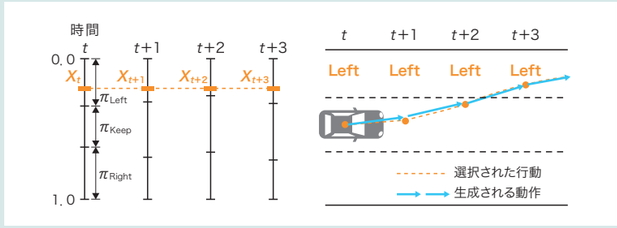

図2．開発した行動探索手法
行動選択に用いる補助変数Xの値を一定期間にわたって固定することで，様々な動作を生成できるようになります。

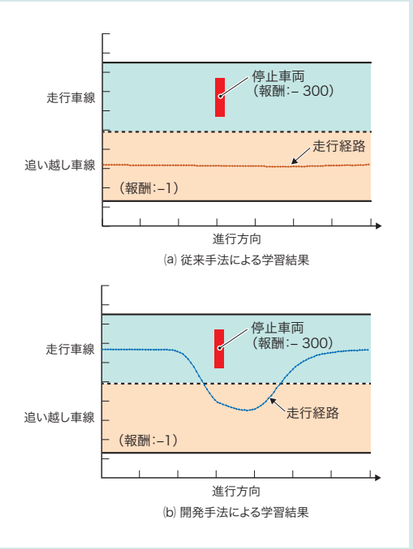

図3．停止した車両を回避する走行の学習結果
開発手法では，100 ms周期でも適切な走行車線の選択を学習できるようになりました。

背景

自動車の自律走行に向けて，従来は車両周辺の状況に応じた走行パターンを，開発者が直接設計するルールベースの手法が開発されてきました。しかし，一般道のように複雑な環境下での自律走行は，従来のルールベースの手法で対応することが困難です。このため，近年，このような複雑な状況にも対応可能な深層強化学習手法を用いた自律走行計画技術が検討され始めています[(1), (2)]。

強化学習は，試行錯誤によって得られた経験に基づいて，より適切な行動を自律的に学習する機械学習手法で，近年これに畳み込みニューラルネットワークなどの深層学習を取り入れた深層強化学習手法が盛んに検討されています。しかし，従来の深層強化学習手法は，行動を決定する周期が長くないと，十分な学習ができないという問題がありました。このため，自動車向けの画像認識で一般的な実行周期100 msに比べて長い，400 ms[(1)]や3.5 s[(2)]といった行動決定周期の下で，深層強化学習が行われてきました。行動を決定する周期が長いと，二つの連続する行動決定の間に想定外の事象が起こる可能性が高くなったり，想定外の事象が起きた場合の対応が遅くなったりすることから，行動決定周期を短くすることは自動車の自律走行計画技術を確立する上で，キーとなります。

従来の行動探索手法の課題

強化学習では，試行錯誤を行って，その学習過程で行動の探索を行います。走行車線の選択における行動探索の概要を，**図1**に示します。行動探索では，一定周期ごとに各時間ステップでの行動（走行する車線）を，各行動の選択確率πに従って選択します。具体的には図1左側に示すように，［0.0，1.0］上の一様分布から各時間ステップで独立に抽出された値Xを，補助変数として用いることで行います。各行動のπは，車両周辺の状況に基づくニューラルネットワークにより算出される値で，学習が進むに連れて適切な行動ほどより大きくなるように変化しますが，学習の初期段階では，図1の両矢印の長さで示すように，左側車線，同車線，及び右側車線の選択確率π_{Left}，π_{Keep}，及びπ_{Right}はほぼ同じ大きさの値となっています。

このような行動探索手法の場合には，行動を決定する周期が短くなると，図1右側に示すように，連続する各時間ステップで選択された，オレンジ色の破線の各行動が短くつなぎ合わされ，結果として，水色の矢印で示すように車線維持と大差がない動作ばかりが発生します。したがって，車線変更のような動作が発生する確率は，非常に低くなります。

強化学習は，様々な動作を試行錯誤し，得られた結果の報酬に基づいて適切な行動を選択する学習手法であるため，行動探索において変化に富んだ動作の生成ができなければ，適切な行動を学習することが困難になります。

開発した行動探索手法

そこで，東芝は，行動探索において行動選択の一貫性を考慮する独自の強化学習手法を開発しました。具体的には，**図2**左側に示すように，行動探索において従来は各時間ステップで独立に抽出していた補助変数Xの値を，開発手法では連続する幾つかの時間ステップにわたって一定の値となるようにします。これにより，連続する時間ステップで同じ行動がより選択されやすくなり，図2右側に示すように車線変更を実現する動作も十分に発生するようになります。

行動探索において行動選択の一貫性を考慮することで，従来は困難であった変化に富んだ動作の生成が行われる結果，適切な行動の学習が可能になります。

開発手法の性能検証

開発手法の効果を検証するため，従来手法と開発手法についてそれぞれ100 ms周期で走行車線を選択する学習を行い，性能を比較しました。学習は，運転シミュレーターを用いて，片側2車線の道路で走行車線上に停止した車両を回避して走行するタスクについて行いました。また，停止車両だけでなく，進路変更における道路交通法の規定を考慮して追い越し車線にも，負の報酬を設定しました（**図3**）。

1千万ステップ分の学習によって得られた，両者の走行経路の比較を図3に示します。従来手法では常に追い越し車線を走行する行動（オレンジ色の走行経路）が学習されたのに対して，開発手法では停止車両を追い越すときにだけ追い越し車線を走行する，適切な車線選択（青色の走行経路）を学習できました。

今後の展望

ここでは，深層強化学習手法だけによる自律走行計画技術について述べましたが，試行錯誤による学習には膨大な時間を要します。今後，自明な走行パターンについては深層強化学習手法にルールベースの手法を組み合わせるといった手法も，検討していきます。

文　献

(1) Mukadam, M. et al. "Tactical Decision Making for Lane Changing with Deep Reinforcement Learning". 2017 NIPS Workshop on Machine Learning for Intelligent Transportation Systems. Long Beach, CA, 2017-12, NIPS. 2017. <https://openreview.net/pdf?id=HylddmUAZ>, (accessed 2020-08-24).
(2) Mirchevska, B. et al. "High-level Decision Making for Safe and Reasonable Autonomous Lane Changing using Reinforcement Learning". 2018 21st International Conference on Intelligent Transportation Systems (ITSC). Maui, HI, 2018-11, IEEE. 2018, p.2156–2162.

野中　亮助
研究開発センター 知能化システム研究所 メディアAIラボラトリー