The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

深層強化学習による施工機械の経路生成

Route Generation of Construction Machine by Deep Reinforcement Learning

田邉峻也 Shunya Tanabe　孫澤源 Zeyuan Sun　中谷優之 Masayuki Nakatani　内村裕 Yutaka Uchimura

芝浦工業大学
Shibaura institute of technology

Recently, research on artificial intelligence has been progressing in various fields. In some of the Atari 2600games, the AI player has scored higher than the skilled human players by using deep reinforcement learning techniques. In this paper, autonomous ground leveling work by a bulldozer is targeted, which are expected to optimize action of the bulldozer. In previous work, we implemented deep Q learning method by giving images of simulator as the input data for the network. However, when learning the image using the convolution layer as input using deep reinforcement learning, it requires large computational cost for learning process. This research aims to reduce the computational cost by giving smaller order of input data. This paper describes the comparison results in different order of input data. Transition of the learning sequence is also evaluated.

1. 緒言

近年，自動車の自動運転，囲碁でプロ棋士に勝利する [David 16] など様々な分野で人工知能の研究が進んでいる．Atari2600 のゲーム [Bellemare 13] をルールの教示無しに画面の状態とスコアのみからコンピュータが繰り返し学習を行うことで，人間よりもスコアを上回った結果が報告されている [Mnih 15]．同論文では Deep Q Network（DQN）と呼ばれる強化学習の手法が用いられている．DQN は，環境の状況からエージェントが採るべき行動を選択し，報酬によってより良い行動の選択をする Q 学習法 [Riedmiller 09] と，画像処理の分野において大きな成果 [Krizhevsky 12] を上げている深層学習を組み合わせた手法である．さらに発展させた手法として Double Q learning[Hasselt 15] や Dueling Double DQN[Ziyu 16] など様々な手法によって人工知能の性能は向上している．

このような人工知能の応用分野として，本研究では建設作業の自動化を対象として行っている。建設作業のなかでも不整地の整地作業は，切り土，盛り土の状況に応じた運転経路の判断が必要なため，最終的な整地形状（出来型）に到達するための逐次の運転動作は，囲碁における大局観に通じるものがある。よって本研究では，不整地をブルドーザのような施工機械で整地する作業の自律化を想定し，DQN の手法を用いることによりブルドーザが自律的に最適な運転経路を生成することを目指す．本研究では，不整地をブルドーザのような施工機械で整地する作業の自律化を想定し，DQN の手法を用いることによりブルドーザが自律的に最適な運転経路を生成することを目指す．しかし，DQN による学習においては，ニューラルネットワークの入力，層数，各層のニューロン数，学習率など様々なハイパーパラメータの調整が必要であり，正しい値を設定しないと学習時間が長くなることや，学習がうまく行かないことがあるためハイパーパラメータ調整は重要である．ハイパーパラメータの調整には試行錯誤を重ねていく必要があるが，畳み込み層を利用した画像を入力として学習する場合，マルチコアの CPU や GPU を使っても学習の過程に非常に長い時間を要することが課題である．これまでに筆者らは，DQN を整地シミュレーションに適用して経路計画する研究を行ったが [孫 17, 中谷 17]，ネットワークへの入力が画面データのため，入力サイズが大きく計算負荷が高くなることが課題であった。そこで本研究では入力サイズを小さくすることで，ニューラルネットワークの規模を縮減し，ニューラルネットワークにおける計算コストを減らすことで学習の高速化を図ることを目的とする．

連絡先: 内村 裕，芝浦工業大学，東京都江東区豊洲 3-7-5，uchimura@shibaura-it.ac.jp

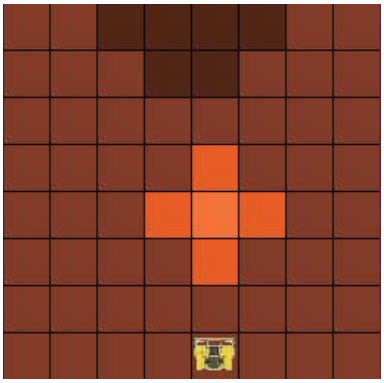


図 1: シミュレーション概要

具体的には，ネットワークへの入力を画像データの代わりに，ブルドーザや土砂の座標データもしくは，ブルドーザと土砂との相対座標とし，入力のサイズを小さくした場合の学習を行い，経路の最適性と高速化について検証する．

2. 問題設定

土砂の整地作業を単純化してシミュレートするために図 1 に示すように整地エリアを 8 × 8 のグリッド上に設定した．エージェントであるブルドーザで土砂山を移動し，図 2 のように地面の凹凸を無くし整地を行うことを目的とする．本シミュレーションでは図 3 のように土砂の高さを 6 段階で表現し，明度が高いと土砂が高く，暗い色ほど土砂が低いことを表している．ブルドーザは上下左右の 4 方向に動き，土砂を押すことがで

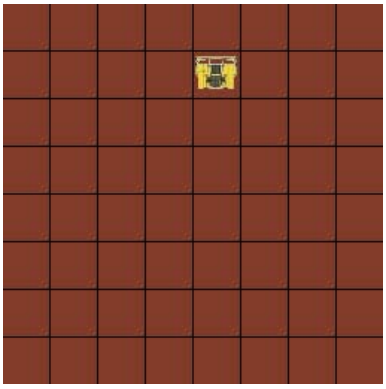


図 2: 整地終了時

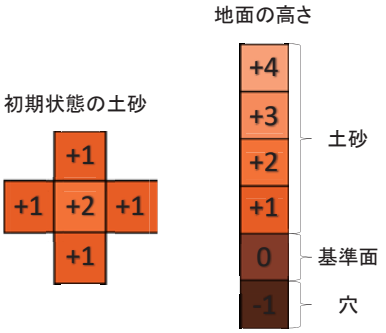


図 3: 土砂と穴の概要

きるが，引くことはできない．基準となる地面の高さを 0 としたとき，高さ −1 の地点の穴に高さ 1 以上の土砂を移動させることで整地を行う．また，図 6 のように，各エピソードにおいて 6 パターンの初期位置をランダムに与えた．報酬の設定は，土砂の高さを −1 の地点に高さ 1 以上の土砂を近づけて移動させることで +1，土砂を 1 エリア整地することで+10，すべてのエリアを整地することで+50 とした．すべての穴が整地されるか，ブルドーザの行動回数が 100 回に達した場合に 1 エピソードの終了とし，エピソードを繰り返し行うことで Q ネットワークの学習を行った．

ネットワークに与える入力の違いによる学習結果の比較のために以下の 2 つの異なる入力データを用意した．

- ケース 1 : ブルドーザの，土砂山・穴の座標値を入力とした場合
- ケース 2 : ブルドーザの周囲のセルの地面の高さを入力とした場合

ケース 2 では入力をシミュレータ画面（8 × 8）より大きなサイズ（15 × 15）のセルとしているが画面外のセルの値は 0 とした。ケース 1 とケース 2 のネットワークの各パラメータを表 1 に示す．中間層に 3 層の全結合層を用いており，各層のニューロン数は 1000，250，50 とした。また，ブルドーザの前後左右方向の移動の 4 つのアクションを出力とした．

学習過程では，割引率 γ を 0.99 に設定した．ϵ-greedy 法によるランダム率は，学習開始時では 100%とし，その後，学習回数が増えるに従って徐々に減らし，学習回数が 100 万回以降は 10%とした．また，学習データを保存するメモリサイズ

表 1: ケース 1 とケース 2 のパラメータ

パラメータ	ケース 1	ケース 2
入力サイズ	38	15 × 15
出力数	4	4
中間層	3 層	3 層
中間層の種類	全結合層	全結合層
ミニバッチ数	50	50
学習率	1.0×10^{-5}	1.0×10^{-5}
割引率	0.99	0.99

表 2: 文献 [孫 17] のネットワークのパラメータ

パラメータ	文献 [孫 17]
入力サイズ	40 × 40 × 4
出力数	4
中間層	7 層（畳み込み 3 層，プーリング 3 層，全結合 1 層）
ミニバッチ数	50
学習率	1.0×10^{-5}
割引率	0.99

は 10 万エピソード分のリングバッファとした．本研究では，Double Q learning を適用し，教師信号を出力するニューラルネットワークはランダムに初期化すると同時に，1 万回の学習ごとに学習済みのニューラルネットワークで上書きした．シミュレーションの 1 エピソードの最大行動回数は 100 回に設定した．また学習回数が 10 万回間隔で学習したネットワークを保存し，ランダム率を 0%としたシミュレーションを行い，行動回数と整地した穴の数を記録した．なお，シミュレータの作成には Pygame を使用し，Double Q learning の実装においては，TensorFlow をフレームワークとして用いた。また，GPU として GeForce GTX 1070（NVIDIA 製）を使用した．

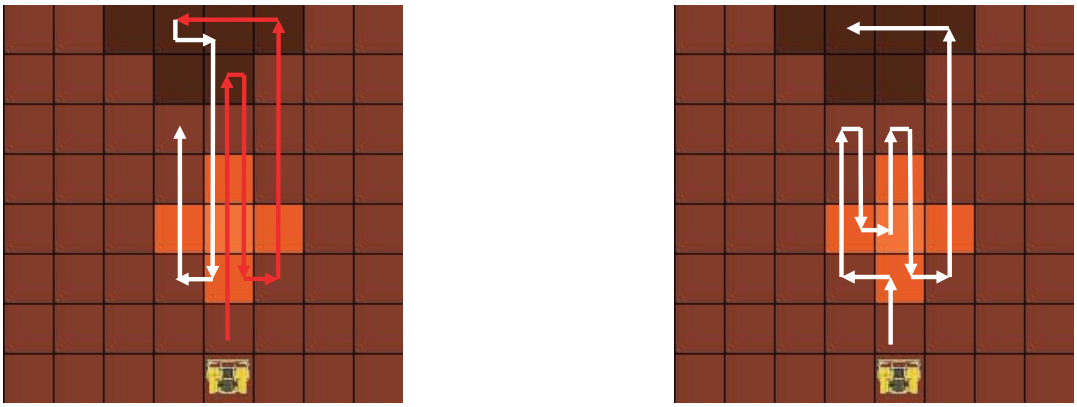
さらに，ニューラルネットワークに与える入力サイズによる計算コストの比較のために，文献 [孫 17] における画像データを入力とした場合の学習も行った．文献 [孫 17] では，1 グリッドを 5 × 5 ピクセルで構成した 8 × 8 のグリッドのシミュレータの画面（40 × 40 ピクセル，8bit 階調）のピクセルデータを 1 画面として，過去 3 画面を含めて 4 画面分を入力としている．同ネットワークに使用したパラメータを表 2 に示す．中間層には畳み込み層が 3 層，プーリング層が 3 層，全結合層が 1 層の合計 7 層をニューラルネットワークに与えられる．しかし，ニューラルネットワークの規模が大きいため，学習時間が長くなるのが問題点である．

3. シミュレーションによる学習結果と比較

3.1 入力データの相違による学習結果

入力層に与える入力データとしてをブルドーザの，土砂，穴の座標値とした場合（ケース 1）と，ブルドーザの周囲のセル（15 × 15）の地面高さを入力とした場合（ケース 2）のシミュレーションによる学習結果を図 5 に示す．グラフ横軸が学習回数，縦軸が各初期位置の穴をすべて整地するまでの行動回数の平均を示している．本結果に示すようにケース 1 ではブルドーザの行動回数が減少せず，学習回数 1,300 万回の試行を行っても穴を全て埋めるには至らず，学習による経路生成が不調に終わった．

ケース 2 では，学習を重ねるごとにブルドーザが全ての穴



(ケース 2)　　(文献 [孫 17])

図 4: 生成経路の比較

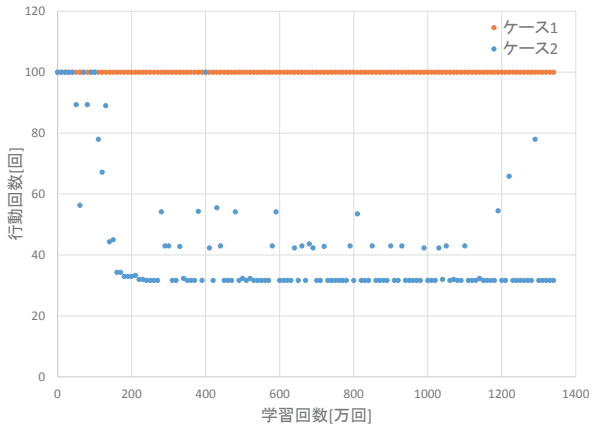


図 5: ケース 1 とケース 2 の学習結果の比較

を整地するまでの行動回数が減少していることがわかる．また穴や土砂の初期位置の異なるパターンにおいても整地を完了する経路の最適化することができた．また，学習回数が 240 万回のときに保存したニューラルネットワークを用いてランダム率 0% でブルドーザをシミュレーション上で動かしたときの 6 パターンの各初期位置の平均行動回数が 31.6 回となった．その後，学習回数 1,300 万回まで学習を行ったが，ブルドーザは 31.6 回より行動回数が少なくなることはなかった。

検証結果が示すように，ケース 2 は学習が進むにつれ整地を完了するまでの行動回数を減少したが，ケース 1 は最大行動回数の 100 回以内に全ての穴の整地を行うことができなかった．また，学習回数を増やしてもすべての穴を整地するまでのブルドーザの行動回数は減少することはなかった．このとき，エピソード開始時は，ブルドーザは上方への行動を数回選択し，いくつかの穴を整地したが，その後全く整地に至る動作を行わなかった．この結果から，ケース 1 で与えられた入力データでは行動価値関数の関数近似が上手く行われていないと考えられる。ケース 1 はケース 2 と比べて入力サイズが少ないことから情報量が少ないこと，入力データが座標データのスカラーの数値のみであり，ケース 2 のような 2 次元配列としての相関がないため学習がうまく進まなかったと推測できる。

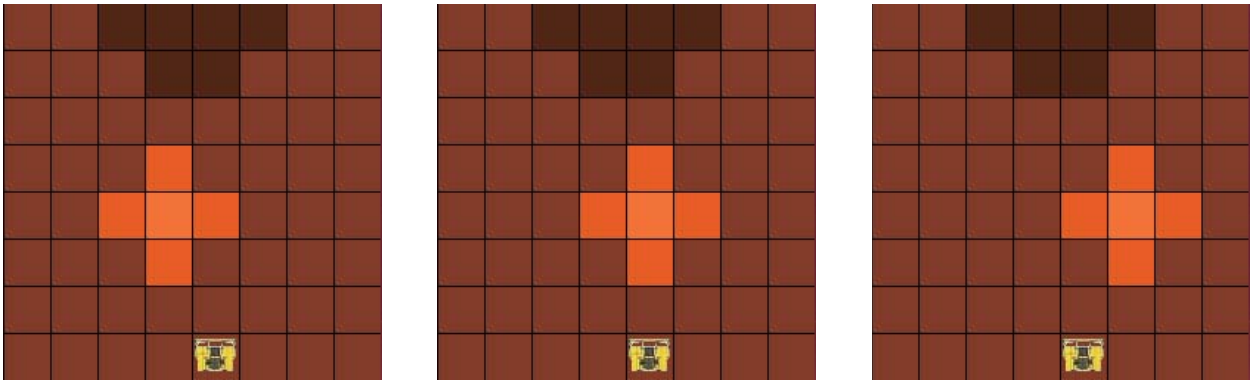
表 3: ケース 2 と文献 [孫 17] の比較

パラメータ	学習時間 [s]	最小行動回数 [回]
ケース 2	5320	31.7
文献 [孫 17]	12300	29.0

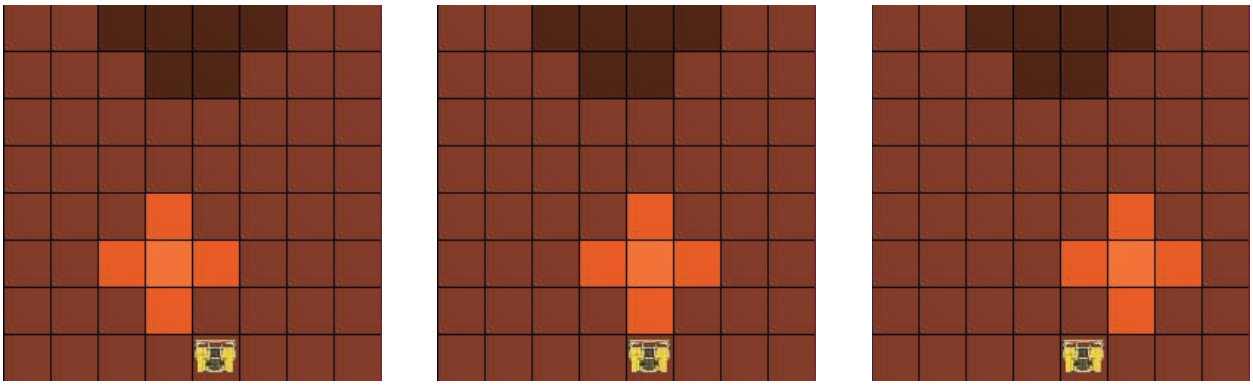
3.2　ケース 2 と文献 [孫 17] の比較

ブルドーザの周囲のセルの地面高さを入力としたケース 2 と，文献 [孫 17] の学習結果の比較を図 7 に示す．縦軸に初期位置の異なる穴をすべて整地するまでのエージェントの行動回数の平均を示し，横軸に学習時間を示している．本結果が示すように，文献 [孫 17] が収束するまでの時間が約 5 時間かかるに対して，ケース 2 とした場合では約 2 時間で収束することがわかる．また，ケース 2 と文献 [孫 17] を学習回数 100 万回にかかる時間とすべての穴を埋めるまでのブルドーザの平均行動回数を比較したものを表 3 に示す．これらの結果が示すように，ケース 2 では文献 [孫 17] と比べて入力サイズを少なくすることでニューラルネットワークの規模が小さくなり，学習時間が約 2 分の 1 に短縮できたことがわかる．しかし，文献 [孫 17] のブルドーザの最小行動回数が 29.0 回に対して，ケース 2 では 31.7 回と行動回数が多くなっていた．また，同じ初期位置のブルドーザの経路をケース 2 と文献 [孫 17] で比較したものを図 4 に示す．矢印はすべての穴を整地するまでのブルドーザの軌跡であり，ブルドーザが 1 マス移動することで行動回数が増えていく．この初期位置の場合ではブルドーザの行動回数は，ケース 2 としたとき 28 回，文献 [孫 17] では 22 回となった．ケース 2 を入力として学習された経路では，最初に 4 つの土砂を運んで穴の整地を行っているが，最後の 2 つの土砂を 1 つずつ運ぶことにより，ブルドーザの行動回数が増えていることがわかる．文献 [孫 17] の経路は，最初に 1 つずつ土砂を運んでいるが，これ以上行動回数を減らすことができない最も効率の良い経路であった．

本研究で適用した手法 (Double DQN) では，現在の状態においてエージェントが最も価値の高い行動を選択するが，人間と違い現在の状態から未来の状態を予測することができない問題点があり，文献 [孫 17] に比べても効率が悪い経路を選択している．文献 [孫 17] と違い，入力数が少ないことや，ニューラルネットワークの規模，畳み込み層の有無などが影響して局所解に陥ってしまった結果，最短経路に到達できなかったと考えられる．



(a) (b) (c)



(d) (e) (f)

図 6: 初期位置

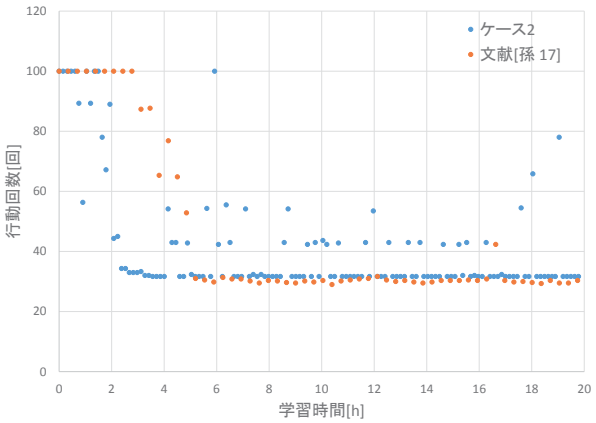


図 7: ケース 2 と文献 [孫 17] の学習結果の比較

4. 結言

本研究では，ブルドーザによる整地作業の自律化を目的に，ニューラルネットワークの入力データとして，ブルドーザや土砂の座標データ，もしくはブルドーザ周囲のセルの状態を与えて学習結果の比較を行った．ブルドーザ周囲のセルの状態を与えることで入力サイズを小さくすることで，ネットワークの規模も縮減し，学習の高速化を実現した．今後は経路をより最適化することを目指すと同時に，より現実に近い問題設定における学習の高速化を図る予定である．

参考文献

[David 16] David, S., Aja, Huang. ”Mastering the game of Go with deep neural networks and tree search” Nature 529 484/489, (2016)

[Bellemare 13] Bellemare, M. G. Naddaf, Y. Veness, J. Bowling, M. :”The arcade learning environment: An evaluation platform for general agents” Journal of Artificial Intelligence Research 47 253-279 (2013)

[Mnih 15] Mnih, V., Kavukcuoglu, K. and Silver, D. ”Human level control through deep reinforcement learning” Nature 518 pp.529-533, 2015.

[Riedmiller 09] Riedmiller, M. Gabel, T. Hafner, R. :”Reinforcement learning for robot soccer” Auton Robot (2009) 27: 55-73 DOI:10.1007/s10514-009-9120-4 (2009)

[Krizhevsky 12] Krizhevsky, A. Sutskever, I. Hinton, G. :”ImageNet classification with deep convolutional neural networks” Adv. Neural Inf. Process. Syst.25, 1106-1114 (2012)

[Hasselt 15] Hasselt, H., Guez, A. and Silver, D.”Deep Reinforcement Learning with Double Q-learning” arXiv:1509.06461v3 cs.LG (2015)

[Ziyu 16] Ziyu, W., Tom, S., Matteo, H., Hado, V,H., Marc, L., Nando, F. ”Dueling Network Architectures for Deep Reinforcement Learning” arXiv:1511.06581v3 cs.LG (2016)

[孫 17] 孫澤源, 中谷優之, 内村 裕, ”深層強化学習による整地作業の自律制御”, ロボット学会学術講演会論文集, 3I2-04, (2017)

[中谷 17] 中谷優之, 孫澤源, 内村 裕, ”深層強化学習による最適経路学習”, 第 60 回自動制御連合講演会, SuD1-4, (2017)