実時間学習に向けた知識転移による二輪ロボットの学習時間の削減*

小谷 直樹[†]

Reduction of Learning Time for Differential-Wheeled Mobile Robots by Knowledge Transfer for Real-Time Learning*

Naoki KOTANI[†]

1. はじめに

強化学習[1]は、エージェントが環境から得られる報酬を頼りに自律的に行動を獲得する手法である。これまでに、ロボットの動作生成への活用を目指した研究等、さまざまな領域への展開が模索されている。強化学習をロボットの動作生成に用いる利点は、設計者がロボットの具体的な動作を設計しなくてもよい点である。反面、強化学習は、エージェントが試行錯誤を通して、環境に適した行動を自ら学習するため、一般的に、多くの試行回数を必要とする問題がある。とくに、エージェントが複数のタスクを扱うような場合、未知のタスクに遭遇するたびに学習し直すことは効率が悪い。

この問題に対して、過去に獲得した知識や、あらかじめ与えられた知識を応用することによって、学習にかかる試行回数を抑制することが考えられる。このような考え方を実現する方法として転移学習があり、近年の機械学習分野の研究でも注目されている話題の一つである。この転移学習を強化学習に適用することで、学習能力の向上を目指した研究事例が報告されている[2–4]。

筆者らは、これまでに、遺伝的アルゴリズムの交叉・淘汰・突然変異の考え方を取り入れた強化学習手法を提案してきた[5]。そして、数値シミュレーション上の多リンク型ロボットアームによる経路獲得問題を対象として、学習能力の向上とシミュレーションに要する時間を大幅に削減できることを示し、提案手法の有効性を明らかにしてきた。しかしながら、これまでのシミュレーションは、摩擦やすべり等の物理的な作用を反映しておらず、現実環境下での学習を想定した場合の有効性については明らかではなかった。この点において、実ロボットを用いて実験することも考えられるが、実ロボットを用いた実験には環境ノイズなど、さまざまな要因が含まれるた

Fig. 1 Learning model

め、学習性能のみを評価することは難しい。そこで、本研究は、三次元物理シミュレータ上で現実に近い環境を再現し、その中でロボットによる行動学習を行うことで、実時間での学習に向けた提案手法の有効性を明らかにする。

2. 学習手法

本研究で用いる学習手法のモデルをFig. 1に示す。この学習手法のモデルは、Actor-Critic[1]とFuzzy ART (Adaptive Resonance Theory)[6]から構成されている。ARTは、観測した状態をカテゴリーという離散状態に分類するニューラルネットワークの一つである。これに、遺伝的アルゴリズムの交叉、突然変異、淘汰の考え方を導入することによって、エージェントに知識転移、局所行動からの脱出、知識整理の機能を与えている。

具体的には、交叉は、エージェントの観測状態が未知であるとき、その未知状態に対して類似度の高い既知の状態が持つ政策を初期値として転移する。これによって、政策の初期値をデフォルト値や乱数を初期値として学習する場合に比べて、タスク達成の可能性が高い政策を利用できるようになるため、学習時間の削減を期待できる。一方、知識の転移は、獲得済みの知識を利用するため、知識の転移を繰り返すと、局所行動に陥ることも考えられる。これを回避する手段として突然変異を用いることで、ある確率で政策にランダム性を与えることが可能で

* 原稿受付 2020年8月28日

[†] 大阪工業大学 情報科学部 Faculty of Information Science and Technology, Osaka Institute of Technology; 1-79-1 Kitayama, Hirakata city, Osaka 573-0196, JAPAN

Key Words: reinforcement learning, transfer learning, knowledge, multi task, differential-wheeled mobile robots.

ある。また、淘汰は、状態価値を基準として必要な知識と不要な知識を分類し、不要な知識を削除する。これによってタスク達成に寄与しないと思われる価値の低い知識が消え、知識の転移によって逆に学習性能が低下する negative transfer の発生を抑えている。加えて不要な知識を削減することは、全体の計算量を抑制する効果もある。以下に、Fig. 1 の各学習ステップの概要を示す。

- step 1 エージェントが環境から状態 $s_t \in S$ を観測。
- step 2 観測した状態 s_t を Fuzzy ART によってカテゴリ c_t に分類。
- step 3 該当カテゴリの政策 $\pi_t(c, a)$ に従って行動を選択。
- step 4 行動 $a_t \in A(c_t)$ に従って、環境に対して行動。
- step 5 行動の結果を TD 誤差によって評価、政策を更新。
- step 6 決められたルールに従って、不要な知識を消去。
- step 7 step 1 へ戻り、繰り返す。

3. 二輪ロボットの経路学習実験

3.1 実験目的

これまででは、シミュレーションの実行時間削減の観点から、知識の転移や整理の有効性を評価してきた。一方で、ロボットの動作時間や物理的な作用を考慮しない数値計算であったため、実環境下でのロボットを使った学習実験を想定したときに、学習時間をどの程度削減する効果があるか、明らかではなかった。実環境下での実験は外部環境やセンサノイズなどがあり、実施の難しさもあるため、今回は三次元物理シミュレータを用いた。

3.2 問題設定

本実験の環境設定を Fig. 2 に示す。この実験は、四方を長さ 500 mm の壁に囲まれた平面内の中心に白丸で示された二輪ロボットが、右上隅の半径 150 mm の扇状のゴール領域に到達するための行動を学習することが目的である。ゴール領域へのロボットの到達判定は、二輪ロボットの中心がゴール領域に入ったときとし、このときのみ +10 の正の報酬をエージェントに与える。また、このタスクは、知識転移の効果を検証するため、正方形で示した一辺 50 mm の四角柱の障害物を設置する場所 (x_{ob}, z_{ob}) を変更することで、異なるタスクを表現する。障害物の設置位置と対応するタスクの番号を Table 1 に示す。ただし、Task 1 では、障害物を設置しない。エージェントは、Task 1 から Task 10 までを順に学習する。知識転移は、500 試行目以降 (Task 2 以降) から行う。これによりエージェントは、Task 2 以降、過去のタスクで獲得した知識を利用しながら学習する。学習の試行回数は、1 タスクにつき 500 試行 $\times$ 10 タスクの計 5000 試行とした。1 試行におけるエージェントの行動可能回数は 20 ステップ、1 ステップ = 512 ms とした。

状態空間は、ロボットの位置と姿勢 (x, z, θ) 、カメラから得られるゴールの色情報、距離センサ $\times 6$ 、の計 10 次元で構築した。Actor-Critic の政策は、左右のモータの回転量を平均 0、標準偏差 10/3 を初期値とするガウス分

Fig. 2 Learning environment (ex. task 7)

Table 1 Positions of an obstacle in each task

Task No.	x_{ob}	z_{ob}
Task 1	-	-
Task 2	0 mm	-200 mm
Task 3	100 mm	-200 mm
Task 4	100 mm	-100 mm
Task 5	0 mm	-100 mm
Task 6	0 mm	-150 mm
Task 7	50 mm	-100 mm
Task 8	50 mm	-150 mm
Task 9	50 mm	-200 mm
Task 10	100 mm	-150 mm

Fig. 3 Path planning problem for a differential-wheeled mobile robot (ex. task 3)

布で表現し、勾配法 [7] によって政策の更新を行う。政策の学習率と価値の学習率は共に 0.2、割引率は 0.9 とした。

実験には、ロボットシミュレータソフトウェアである Webots を用いた。ロボットの一例として、直径約 70 mm の二輪ロボットである e-puck を用いた。実際のシミュレーション実験の様子を Fig. 3 に示す。中央の円筒形状のものがロボットである。本実験では、知識転移の有無による学習結果について比較する。

3.3 実験結果

知識転移ありとなしの場合、それぞれについて、10 個の異なる乱数パターンを与えてシミュレーションを実施した。エージェントの学習を評価する学習曲線は、ある 100 試行中のゴール到達回数の移動平均である。その 10 個の移動平均の平均値をとった学習曲線を Fig. 4 に示す。

Fig. 4 Comparison of success rate

図中の丸で囲んだ数字は、Table 1 のタスク番号を表す。

Task 1 は知識転移を行わないため、学習曲線の推移は一致している。Task 2 以降の学習曲線を比較すると、知識転移ありの場合、いずれのタスクにおいても高い成功率を示している。とくに、Task 3, 4, 5, 7, 8, 10 の学習曲線の差が顕著である。Task 3, 5, 8 においては、タスクが切り替わった直後の成功率の低下が大きく抑えられており、知識転移によって、過去のタスクで得た知識を有効に活用することによって、タスクの変化に早期に対応できていることがわかる。一方、知識転移なしの場合、多くの場合においてタスクが変化すると成功率が 10 % 程度低下しており、とくに、Task 4, 10 においては、タスクが終了する段階になっても 80 % 台の成功率であった。

つぎに、各タスクにおいて、学習曲線が 95 % を超えるまでの時間を Table 2 に示す。Task 1 は、共通であるため除外している。表記が 0 である個所は、タスクが変わっても成功率が 95 % を下回らなかったことを表す。また、知識転移なしで、Task 4, 10 においては 1 度も成功率が 95 % を超えることはなかった。結果から、知識転移は、これを使わない場合に比べて、学習時間が 0 であるタスクの数が 2 から 5 へ増えており、これらだけで、100 分以上の学習時間を削減した。また、Task 2, 7 においても、95 % の成功率に到達するまでの時間が半分以下となっている。さらに、知識転移は、知識転移がない場合では 500 試行で十分に学習できなかった Task 4, 10 においても、25 分程度で学習できた。

4. おわりに

本研究は、実時間学習への提案手法の適用に向けて、三次元物理シミュレータ上の二輪ロボットの経路獲得問題を対象に有効性を検証した。その結果、知識転移は、タスクが切り替わった直後において高い成功率を示しており、環境の変化に早期に対応できることを示した。また、知識転移を用いた方が、安定的に高い成功率を維持していることを確認した。学習時間については、実時間ベースでの学習においても、知識転移によって学習時間

Table 2 Comparison of learning time [min.]

Task No.	with TL	without TL
Task 2	9.7	21.1
Task 3	0	48.3
Task 4	25.0	N/A
Task 5	0	31.9
Task 6	0	0
Task 7	24.1	60.6
Task 8	0	28.7
Task 9	0	0
Task 10	23.4	N/A

を半分以下に削減できることを示唆する結果を得た。

今後の課題として、より規模の大きなタスクやノイズを含んだ環境でのシミュレーションによる検証、実際に二輪ロボットを使い、本提案手法の有効性を明らかにすることが挙げられる。また、今回使用した車輪型のロボットだけでなく、さまざまなロボットを対象として、本手法が広く適用可能であることを明らかにすることが挙げられる。

謝 辞

本研究は、JSPS 科研費 JP16K16118, JP20K11943 の助成を受けたものです。ここに感謝申し上げます。

参考文献

- [1] R. S. Sutton and A. G. Barto (著), 三上, 皆川 (共訳): 強化学習, 森北出版 (2000)
- [2] M. E Taylor and P. Stone: Transfer learning for reinforcement learning domains: A survey; *Journal of Machine Learning Research*, Vol. 10, No. 1, pp. 1633–1685 (2009)
- [3] D. Abel, Y. Jinnai, S. Y. Guo, G. Konidaris and M. Littman: Policy and value transfer in lifelong reinforcement learning; *Proceedings of the 35th International Conference on Machine Learning*, PMLR Vol. 80, pp. 20–29 (2018)
- [4] A. Barreto, D. Borsa, J. Quan, T. Schaul, D. Silver, M. Hessel, D. Mankowitz, A. Zidek and R. Munos: Transfer in deep reinforcement learning using successor features and generalised policy improvement; *Proceedings of the 35th International Conference on Machine Learning*, PMLR Vol. 80, pp. 501–510 (2018)
- [5] 小谷: 学習時間の短縮に向けた状態価値を用いた知識転移手法; 電気学会論文誌 C, Vol. 137, No. 9, pp. 1171–1176 (2017)
- [6] G. A. Carpenter, S. Grossberg and D. B. Rosen: Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system; *Neural Networks*, Vol. 4, pp. 759–771 (1991)
- [7] R. J. Williams: Simple statistical gradient-following algorithms for connectionist reinforcement learning; *Machine Learning*, Vol. 8, pp. 229–256 (1992)