転移学習における心理学モデルを用いた方策再利用手法の検討

高桑 優作[*1]，河野 仁[*2]，温 文[*3]，鈴木 剛[*1]

A Study on Policy Reuse Method Using Psychologically Inspired Model in Transfer Learning

Yusaku TAKAKUWA[*1] Hitoshi KONO[*2]
Wen WEN[*3] and Tsuyoshi SUZUKI [*1]

[*1] Department of Information and Communication Engineering, Tokyo Denki University
Senju-Asahi-chou 5, Adachi-ku, Tokyo, Japan
[*2] Department of Electronics and Mechatronics, Tokyo Polytechnic, Japan
1583 Iiyama, Atugi, Kanagawa, Japan
[*3] Department of Precision Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

This paper describes a policy transfer method of reinforcement learning agent based on spreading activation model of cognitive psychology. A reinforcement learning agent accumulates policies which are learned with various environmental settings, and the agent selects and transfers a policy which is presumed to be optimal based on the situation of the agent while learning. In the existing method, according to the target-task, human evaluated and selected the transfer policy manually. The purpose of this research is to develop automation of policy reuse method. In the proposed method, an undirected graph is generated between policies and a network is constructed between the measures using the undirected graph. Based on the graph structure, the agent adjusts the activated value of the policy while repeating selection, activation, and spread processing, and decides a policy to select probabilistically and transfers. In this study, we compared the total rewards of the proposed method using multiple policies and Deep Q-Network with the learning model generated randomly and verified the usefulness by computer experiment.

Key Words : Cognitive Robotics , Reinforcement Learning , Transfer Learning

1. はじめに

近年，学習能力，認識能力等のような知的能力を有するロボットの実用化が期待されている (1)．未知環境の考慮や，多様に変化する状況に応じて人が予めロボットに制御則を与えておくことは困難であることから，ロボット自身に自律的に学習をさせる強化学習の研究が盛んに進められている (2)．

強化学習 (3) とは，エージェント（以下，学習可能なロボット，システムをエージェントと呼称）に試行錯誤を行わせることで，最適な行動を学習させる手法である．強化学習によって，エージェントは自律的に行動則の獲得が可能になるが，実用的，複雑なタスクの学習には，長時間の学習時間が必要になるという課題点も存在する．この課題点に対しては，予め学習したタスク（以下，Source-task と呼称）の方策を学習予定のタスク（以下，Target-task と呼称）で再利用する学習手法である，転移学習 (4) ~ (6) と呼ばれる手法が検討されている．転移学習により，新たなタスクや環境への適応能力の向上や，学習時間の短縮が可能になる．しかし，既存研究の転移学習においては，予め Target-task の学習に有効であると考えられる Source-task の方策を決定しておく必要があるため，人の手による方策の評価が必要になる．本研究では，強化学習を行うエージェントが，獲得・保存した方策の特徴に着目し，複数の方策を Target-task の学習状況に応じて選択しながら，効果的に学習可能な方策をエージェント自身に決定させる新たな転移学習手法を提案

[*1] 東京電機大学大学院工学研究科情報通信工学専攻（〒120-8551 東京都足立区千住旭町 5）
y.takakuwa@nrl.c.dendai.ac.jp, tszk@mail.dendai.ac.jp
[*2] 東京工芸大学工学部電子機械学科（〒243-0297 神奈川県厚木市飯山 1583）h.kono@em.t-kougei.ac.jp
[*3] 東京大学大学院工学系研究科精密工学専攻（〒113-8656 東京都文京区本郷 7-3-1）wen@robot.u-tokyo.ac.jp

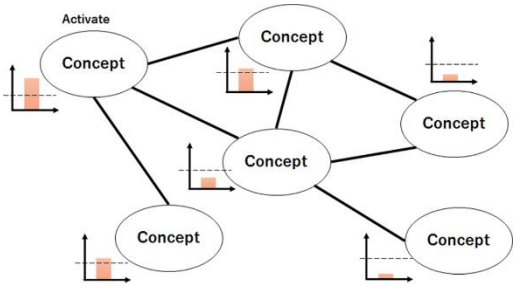

Fig.1　Example of spreading activation theory

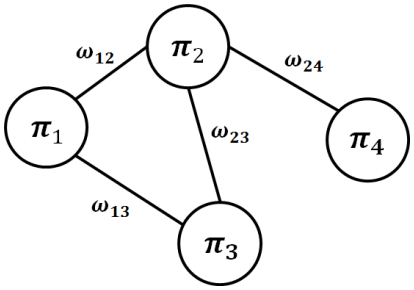

Fig.2　SAP-Net

する．提案手法を検討するにあたり，ヒトは学習した情報を選択，考慮しながら学習しているという認知心理学的知見 [7] を考慮し，心理学モデルである活性化拡散モデル [8] を用いた手法を検討する．

本稿では，活性化拡散モデルを用いた転移学習手法を開発し，計算機実験により提案手法の有用性を検証したので報告する．

2.　研究背景

強化学習で用いられる方策記述法は大きく 2 種類に分類できる．一つは，学習して獲得した行動価値を Q テーブルと呼ばれる look up table に記述する方法である． Q テーブルを用いる場合は，エージェントの状態と行動の組み合わせを全て記述するのが特徴であり，転移学習に利用する際には，テーブルから状態毎に行動価値を参照し再利用が可能になる．しかし，look up table を利用する場合は状態数が増えると行動価値の組み合わせの増大により，使用するテーブルサイズが指数関数的に増加してしまう欠点が存在する．そのため，実環境での複雑なタスクの学習が困難である．

この課題に対して，方策記述法の二つ目として行動価値を関数近似する研究が存在する．関数近似を用いた手法は複数存在しているが，本研究では DQN [9] 等のニューラルネットワーク（以下､NN）を用いた関数近似手法を取り上げる．関数近似により，Q テーブルに比べ多次元タスクの学習が可能となる．ただし，関数近似を行う際の NN の中間層や活性化関数の設定によっては，学習時間の増大[10]や過学習，未学習を引き起こす．

ロボットに対して複雑なタスクの場合には，関数近似手法を用いた転移学習が有効であると考えられるが，学習環境や規模，タスク毎にネットワーク構造を最適化 [11] することは現実的ではない．また，転移学習において，学習を効率化させるための様々な研究 [12]，[13] が進められているが，人手による処理は最小限にすべきである．そこで，本研究では，ランダムにネットワーク構造を設計，学習した方策を複数保存しておき，転移学習システムが確率的に転移する方策の決定を行う新たな転移学習手法を提案する．

3.　活性化拡散モデル

活性化拡散モデルとは，ヒトが獲得した概念同士が脳内でネットワーク構造として保存されていることを前提としたヒトの概念想起（思い出し，再認識等）に関わる心理モデルである．活性化拡散モデルには，概念同士の関連性の強さに応じて概念間の距離が変化する意味的距離と呼ばれる考え方が存在する．概念の活性化は，関連性によって構築されたネットワークを介して行われる．活性化拡散モデルの例を図 1 に示す．図 1 では，活性化された概念から伸びる距離を経由して活性値と呼ばれる値が拡散している様子を示している．本研究では，方策間の関連性に基づきネットワークを構築し，ネットワークを用いてエージェントに転移する方策を決定させる．

4.　提案手法

4・1　前提条件　本節では，提案手法を述べるにあたり，関係用語や前提条件を述べる．強化学習アルゴリズムには，Q 学習を用いる．本研究では，関数近似した NN の構造（以下，近似モデル）を方策として利用する．

4・2　提案手法の流れ　提案手法の転移学習システムは，転移学習前と転移学習中の二つの手続きによって構成されている．

転移学習前の手続きとして，予め学習した複数のSource-taskの方策の特徴を基にカテゴリに分類する．カテゴリとは，本研究においては複数の方策に関連性を計算した集合を指している．生成したカテゴリを用いて方策のネットワークを作成し保存する．この方策で構成されたネットワークを本研究では，Spreading Activation Policy Network（SAP-Net）と呼び，転移学習中の手続きに利用する．図2にSAP-Netの例を示す．

転移学習中の手続きは，選択，活性化，拡散と

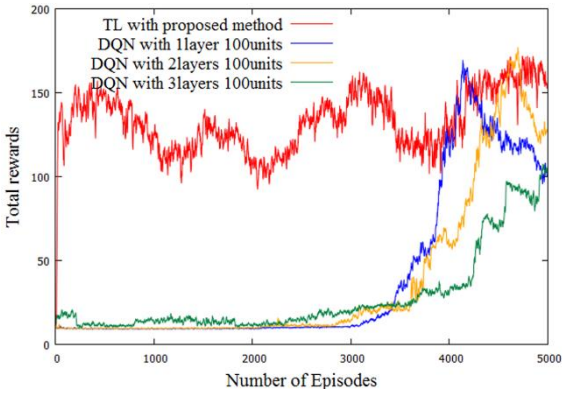

Fig.3　Learning curve with moving average (CartPole task)

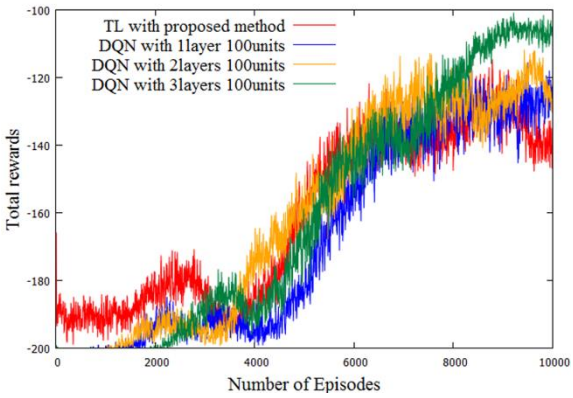

Fig.5　Learning curve with moving average (MountainCar task)

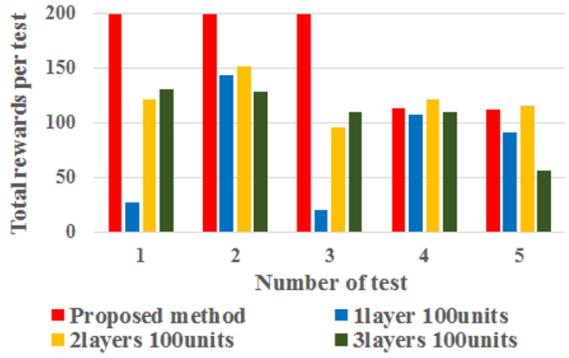

Fig.4　Test task after learning (CartPole task)

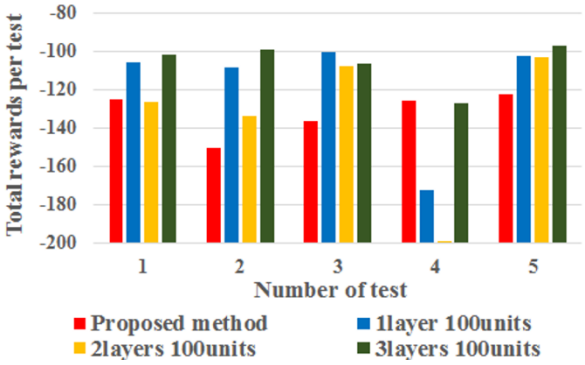

Fig.6　Test task after learning (MountainCar task)

呼ばれる手続きで構成される．選択では，各方策に与えられた活性値と呼ばれるパラメータを基に算出した選択確率を参照して方策を選択する．活性化では，選択された方策を通じて得られた行動毎に算出される行動価値と教師となる値との損失を基に，選択した方策による行動が学習を促進させた場合（以下，正の転移）と促進できなかった場合（以下，負の転移）の判定を行う．この判定結果から，選択した方策の活性値を調整する．拡散では，SAP-Netを基に，選択された方策の活性化を通じて再帰的に各方策の活性値に対して活性化の伝播を行う．これらの転移学習中の処理により，転移学習中に各方策の活性値を変動させる．活性値の変動を利用して，活性値の大きい方策を優先的に学習させるシステムを構築する

5.　実験

5・1　実験目的　本研究では，Source-task の方策を自律的に転移するシステムの有用性を検証するためランダムに中間層のユニット数を設定し学習した近似モデルを複数用いて，提案手法を利用した転移学習（TL：Transfer　Learning）を行う．この新しい転移学習手法と深層強化学習（DQN）の学習効率を比較した．

今回の実験では，Source-taskとTarget-taskを統一した，同一タスクで検証を行った．主に，実験にて正の転移と負の転移を判定するための条件について検証し，方策選択により転移学習の効果が表れていることを確認する．評価として，Episode対総報酬の学習曲線と，学習後の方策を用いてテストを行う．学習曲線は学習の効率を評価しており，テストは学習後にタスクに適した方策を獲得していることを評価している．

5・2　実験設定　実験では，方策転移による近似モデルの重みは固定せずに再学習をさせる．重みの固定をすると近似モデルの入出力ユニット数を Source-task と Target-task 間で合わせる必要がある．また，層の部分転移をする際にも，予め固定する層の指定をする必要がある．この実験設定では，重みの固定をしないことで，複数方策を選択しながら再学習させる新しい転移学習の効果を確認する．提案手法に用いた方策数は 10 とし，カテゴリの分類基準は Source-task で獲得した近似モデルの中間層の層数とした．

5・3　実験結果　CartPole タスクの学習曲線を図 3 に示し，学習後のテストで獲得した総報酬の比較を図 4 に示す．MountainCar タスクの学習曲線を図 5 に，学習後のテスト結果の比較を図 6 に示す．図 3 及び図 5

において，赤の曲線が提案手法を用いた転移学習の結果を表し，青い曲線は 1 層 100 ユニット，橙色の曲線は 2 層 100 ユニット，緑色の曲線は 3 層 100 ユニットの DQN の結果を表している．図 4，図 6 の縦軸は総報酬, 横軸はタスクの実行回数 (Episode) を表している．

図 3 の結果から，提案手法は学習初期の Episode から高い報酬を獲得できていることが確認できる．この結果から提案手法により，複数の方策の中から選択しながら再学習することで，学習の促進がなされていることがわかる．比較対象の 3 種の DQN の学習結果からは，総報酬が低下してしまう現象が見られた．これは，タスクに対して学習モデルの構造が最適化されていないことが原因であると考えられる. 図 4 の CartPole のテスト結果から，比較対象の 3 種の DQN は，提案手法に対して，獲得できる報酬量も低下している．提案手法は，方策を選択しながら再学習が可能になり，学習中に最も優先度合いの高い方策を転移できるため, テスト時にも高頻度で高い報酬の獲得を維持できていることがわかる．図 4 の 4,5 回目のテストで報酬が低下した部分は，再学習時に最適方策の獲得ができなかったものと推測する．この点については，近似モデルの重み固定等を含めた転移の仕方について検討する必要があると考える．以上の結果から総合的に，CartPole において提案手法の有用性を確認することができた．

図 5 から，提案手法は学習初期段階の報酬獲得量が多いものの，学習終了時付近の Episode において，比較対象に比べ，獲得総報酬が減少している．図 6 のテスト結果からも，獲得総報酬が 3 種の DQN に比べ，減少していることが確認できる．この結果は，提案手法において方策の転移判定が行動毎に行われているため，タスク終了時まで正負の転移判定が困難な MountainCar タスクで活用できなかったことが原因であると考えられる．これらの結果に対して，方策の転移判定を Episode 毎に行うことや，一定の行動数毎に行うことを検討することによって，解決可能であると考えられる．また，活性化の判定に利用する指標を，損失以外にも検討することで，より効果的に判定可能な条件を検討できると考える．

以上の結果から，提案手法の利用可能範囲について考察すると，タスク達成条件に報酬付与タイミングが結びつきやすいタスクには適用可能であると考えられる．また，課題としては，現状経験則的に決定しているカテゴリの分類方法において，効果的な学習につながるように定式化された手法の構築が必要になる．

6. おわりに

本稿では，活性化拡散モデルを用いた転移学習として，方策を選択しながら再学習する手法を提案し，計算機実験により検証した．検証結果より，転移学習前の人手による評価，モデル設計によらない同一タスクの転移学習において示唆された．また，課題点として今後は，適切な正の転移，負の転移の判定及びカテゴリの分類基準について検討し，異種タスク，異種エージェントの転移学習において，自律的に Source-task の方策を Target-task に転移可能にする手法について検討していく．加えて，実機ロボットに提案手法を適用するための仕組みについても検討していく．

謝辞

本研究の一部は，JSPS 科研費 JP16K12493 の助成を受けて行われた．ここに謝意を表する．

参 考 文 献

(1) 榊原 伸介，“知能ロボットによる工場自動化と IoT, AI 活用について”, システム制御情報学会誌, Vol.61, No.3(2017), pp.101-106.
(2) 山田 和明，保田 俊行，大倉 和博，“マルチロボットシステムのための状態空間表現を適応的に切替える強化学習”，Vol.84，No.862(2018).
(3) R.S.Sutton ， and A.G.Barto，“強化学習”，森北出版株式会社，(2000).
(4) M.E.Taylor ， ”Transfer in Reinforcement Learning Domain”，Springer，(2009).
(5) S.J.Pan ， and Qiang Yang ， “A Survey on Transfer Learning”，*IEEE Transaction on Knowledge and Data Engineering*，Vol.22，No.10(2010)，pp.1345-1359.
(6) 神嶌 敏弘，“転移学習”，人工知能学会，Vol.25，No.4(2010)，pp.572-580.
(7) 中島 義明・ほか編, “心理学辞典”, 株式会社有斐閣, (2016)，P522-523
(8) A.M.Collins，and E.F.Loftus，“A Spreading –Activation Theory of Semantic Processing”，*Psychological Review*，Vol.82，No.6(1975)，pp.407-428.
(9) V.Minh，*et al*，“Human-level control through deep reinforcement learning”，*Nature*，Vol.518(2015)，pp.529-533.
(10) S. Abe，“Neural Networks and Fuzzy Systems”，Springer，(1997).
(11) A.A.Rusu，*et al*，“Progressive Neural Networks”，Cornell University Library，(2016).
(12) 稲盛 有那，平川 翼，山下 隆義，藤吉 弘亘，柏原 良太，稲葉 正樹，二反田 直己，“事前知識を活用した Memory Reinforcement Learning による行動獲得”，*The 32nd Annual Conference of the Japanese Society for Artificial Intelligence*，(2018)．
(13) Q.Cheng，X.Wang and L.Shen，“Transfer Learning via Linear Multi-variable Mapping under Reinforcement Learning Framework”，*Proceedings of the 36th Chinese Control Conference*，(2017)，pp.8795-8799．