L-005

深層強化学習を用いた経路制御における行動価値関数の変更による性能改善
Performance Improvement of Traffic Engineering Using Deep Reinforcement Learning by Changing an Action Value Function

大石 勲斗† 佐川 勇太† 瓜本 稜† 福島 行信† 樽谷 優弥‡
Isato Oishi Yuta Sagawa Ryo Urimoto Yukinobu Fukushima Yuya Tarutani

1. はじめに

近年，機械学習を用いてネットワークの制御・運用の最適化や自動化を図る Knowledge Defined Networking (KDN) が注目されている[1]．KDN に関する従来研究[2]では，強化学習の一種である深層強化学習[3, 4]を用いた経路制御手法（以下，従来手法と呼ぶ）の有効性が示されている．

従来手法では経路を選択する際に，行動価値関数に基づいて経路の良し悪しを判断している．しかしながら，この行動価値関数が経路制御問題の設定に沿うようには定義されていないため，必ずしも高いネットワーク性能が得られるとは限らないと考えられる．

そこで，本研究では，経路制御問題の設定にあわせて行動価値関数の定義を変更することによってその性能改善を図る．

2. 深層強化学習を用いた経路制御

2.1 深層強化学習

深層強化学習は強化学習の一種であり，強化学習で用いられる方策や行動価値関数といった関数を Deep Neural Network (DNN)で近似する学習法である．図1に深層強化学習の枠組みを示す．深層強化学習では学習のための空間を「環境」, 学習者を「エージェント」とよぶ．深層強化学習では状態，行動，報酬という三つの要素を，環境とエージェントの間でやり取りすることで学習を進める．エージェントは環境の状態を観測し，その状態に対して取るべき行動を決定する．その行動に応じて環境は次の状態へ遷移し，報酬をエージェントへ与える．エージェントの目的は，将来にわたって得られる報酬の総和（累積報酬）の最大化である．

本研究では，深層強化学習の手法として従来研究[2]と同様に Actor-Critic 法を用いる．この手法では，エージェントは「方策」と「行動価値関数」で構成される．方策は，状態から行動を得るための関数である．行動価値関数は，ある状態である行動をとることの価値（累積報酬の期待値）を得るための関数である．いずれの関数も DNN により近似される．

行動価値関数$Q(s_t, a_t)$は式(1)で定義される関数である．

$$\boldsymbol{Q(s_t, a_t) = r_t + \gamma \cdot maxQ(s_{t+1}, a_{t+1})} \quad \textbf{(1)}$$

ここで，s_tは時刻tでの状態，a_tは時刻tでの行動，r_tは時刻tでの報酬，$maxQ(s_{t+1}, a_{t+1})$は時刻$t+1$以降に得られる報酬，γは将来得られる報酬の割引率である．行動価値関数は式(1)のように再帰的な構造を持っているため，将来にわたる累積報酬を求められる．ある環境での行動価値関数がわかれば，エージェントは常に行動価値関数の値が最大となる行動をとることで累積報酬を最大化できる．よって，深層強化学習の目的は行動価値関数を極力正確に近似することであるといえる．

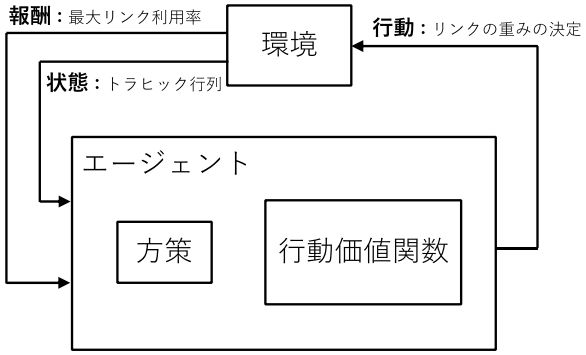

図 1　深層強化学習の枠組み

2.2 経路制御への適用

本研究で取り組む経路制御問題では，送受信ノード間の経路としてリンクの重みの総和が最小となる経路が選択されるという前提の下で，与えられたトラヒック行列(送受信ノード間のトラヒック要求を要素とする行列)に対して，リンク利用率の最大値(最大リンク利用率)を最小化するように各リンクの重みを決定する．各リンクの重みは，与えられるトラヒック行列が変化するたびに個別に決定できる．

この経路制御問題に対して深層強化学習を適用するために，状態をトラヒック行列に，行動を各リンクの重みを決定することに，報酬を最大リンク利用率の反数にそれぞれ対応づける(図 1).

3. 提案手法

3.1 従来手法の問題点

従来手法では，式(1)の行動価値関数$Q(s_t, a_t)$を用いて，現時刻tにおいて状態s_tが与えられたときに行動a_tをとることで得られる累積報酬を算出する．この行動価値関数を本研究で取り組む経路制御問題で用いることは，現在発生しているトラヒック行列に対して各リンクの重みを決定する際に，その決定によって現在の最大リンク利用率がどうなるかのみでなく，将来発生するトラヒック行列に対して，

† 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University
‡ 岡山大学 大学院ヘルスシステム統合科学研究科
Graduate School of Interdisciplinary Science and Engineering in Health Systems, Okayama University

将来に決定される各リンクの重みによって将来の最大リンク利用率がどうなるかをも考慮していることを意味する．

しかし，実際には，将来に発生するトラヒック行列は，現在決定される各リンクの重みとは独立に決まるため，現在決定される各リンクの重みが影響を及ぼすのは現在の最大リンク利用率のみである．したがって，式(1)の行動価値関数を本研究で取り組む経路制御問題で用いることは不適であると考えられる．

3.2 行動価値関数の変更

前節で述べたように，現在決定される各リンクの重みが影響を及ぼすのは現在の最大リンク利用率のみである．そのため，各リンクの重みを決定する際には，現在発生しているトラヒック行列に対して現在の最大リンク利用率のみを最小化するように各リンクの重みを決定すればよい．これを実現するために，行動価値関数を以下のように変更する．

$$Q(s_t, a_t) = r_t \qquad (2)$$

4. 性能評価

4.1 評価モデル

計算機シミュレーションにより提案手法の有効性を評価する．ネットワークモデルとしては，図 2 のスケールフリーネットワーク(14 ノード，平均次数 3)を用いる．トラヒック行列については，重力モデルを用いて，学習用トラヒック行列とテスト用トラヒック行列をそれぞれ 1000 個ずつ生成した． 方策を近似する DNN と行動価値関数を近似する DNN の学習には，ミニバッチを用いた確率的勾配降下法を採用する．

4.2 評価結果

図 3 に提案手法と従来手法の最大リンク利用率を示す．横軸は学習回数を表す．

提案手法は従来手法と比較して，最大リンク利用率を 45%程度改善できている．このことから，提案手法における行動価値関数の変更は有効であるといえる．

提案手法では学習回数が 20,000 回に達するまでの間に最大リンク利用率を大きく低減できている．これは，提案手法では行動価値関数が即時報酬として定義されていることから，少ない学習回数で行動価値関数を正確に近似できたためと考えられる．一方，従来手法では学習回数が 100,000 回に達してもまだ最大リンク利用率は高く保たれている．これは，従来手法では行動価値関数が再帰的に定義されていることから，行動価値関数を近似するために学習の反復が必要であるためと考えられる．

5. むすび

本研究では，深層強化学習を用いた従来の経路制御手法において行動価値関数を変更することにより性能改善を図った．計算機シミュレーションの結果， 45%程度の性能改善を達成できることがわかった．

今後は，様々なパラメータ設定でのより詳細な性能評価を行う予定である．

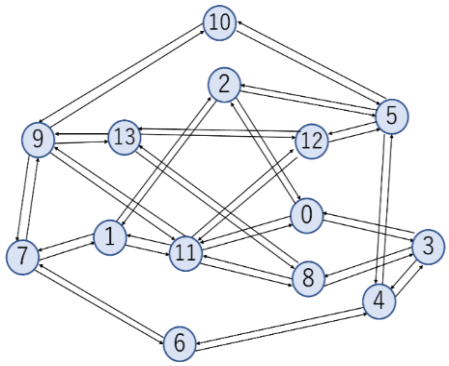

図 2　ネットワークモデル

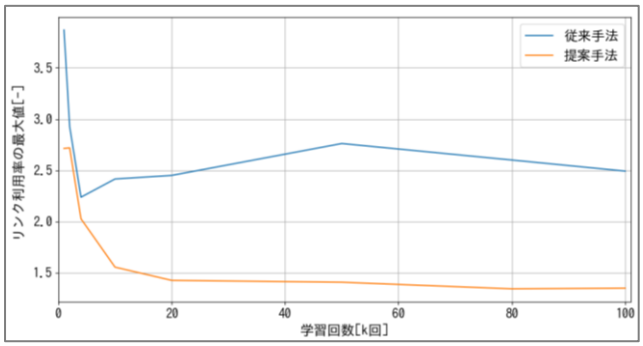

図 3　学習回数と最大リンク利用率の関係

参考文献

[1] A. Mestres, et al.,“Knowledge-Defined Networking,” *ACM SIGCOMM*, vol. 47, no. 3, pp. 2-10, Jul. 2017.

[2] G. Stampa, et al.,“A Deep-Reinforcement Learning Approach for Software-Defined Networking Routing Optimization,” arXiv preprint arXiv:1709.07080 , pp. 1-2, 2017.

[3] Z. Xu, et al.,”Experience-driven Networking: A Deep Reinforcement Learning based Approach,” IEEE INFOCOM, pp. 1-9, 2018.

[4] Y. Li,“Deep reinforcement learning: An overview,” arXiv preprint arXiv:1701.07274, pp. 5-22, 2018.