4P-08

強化学習を用いた巡回セールスマン問題の解法 *

山本　大輔 †
関西大学大学院 †

三木　彰馬 †
関西大学大学院 †

榎原　博之 ‡
関西大学 ‡

1　はじめに

組合せ最適化問題は計算機科学における基本的な問題の 1 つである. 輸送や通信, 製造, インフラ計画など, さまざまな分野における多くの課題は組合せ最適化問題として扱うことができ, 現実社会での応用が期待されている. 典型的な組合せ最適化問題の 1 つに巡回セールスマン問題（TSP : Traveling Salesman Problem）が挙げられる.TSP とは与えられたグラフにおいて, すべての頂点を 1 度だけ通るような巡回路のうちエッジ (辺) の距離の総和を最小とするものを求める問題である. 頂点間の距離が, 直線距離 (ユークリッド距離) である TSP を平面 TSP と呼ぶ.

近年, 深層学習を用いた技術が活発に研究され, 今までは困難であった課題を解決できる可能性がある手法として注目されている. しかし, 深層学習では学習を行うための教師データを用意することが困難であることが多く, 様々な問題を解決するための経験を得ることは難しい. そこで, 自らの行動により学習を行う強化学習を用いることで教師データを用意することが困難な場合であっても様々な経験を学習することができる. このことから, 組合せ最適化問題の解法において強化学習を利用することで, 未知の入力への汎化が得られるのではないかと考えられる.

本研究では平面 TSP に注目し, TSP とその解を画像として扱い, 畳み込みニューラルネットワーク（CNN : Convolutional Neural Network）により最適経路の画像を近似した優良エッジ分布と, ここから得られる各エッジの評価値である優良エッジ値に対して, CNN により得られた解を比較することによって強化学習を適用する手法を提案する. この手法では CNN により得られた解の経路長を各問題例ごとに記憶し, 新たに得られた解を過去の解と比較することによって良い解の経路を選ばれやすく, 悪い解の経路を選ばれにくくなるように学習する.

2　CNN を用いたエッジの評価

2.1　優良エッジ分布

TSP に対して深層学習を適用した手法として点および線の描画によってその問題例と解を画像として表現する手法 [1] がある. ある平面 TSP の問題例について, すべての頂点を描画した頂点画像と, その最適経路を描画した最適経路画像を入力とし, その問題例の最適経路画像を出力するようなモデルを CNN を用いることで近似し, その出力を優良エッジ分布と呼ぶ. 優良エッジ分布の出力例を図 1 に示す.

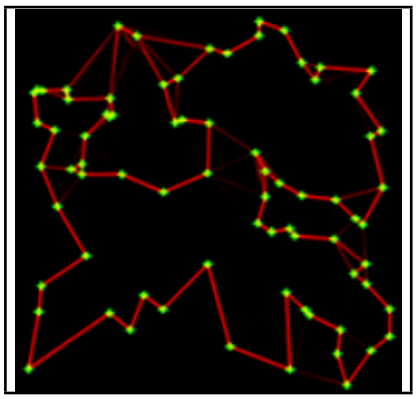

図 1　優良エッジ分布の出力例

2.2　優良エッジ値

学習により優良エッジ分布が得られたとき式 (1) のように各エッジ (i, j) 上における優良エッジ分布の平均を求めることで, そのエッジが最適経路に含まれる尤度を計算する. これを優良エッジ値と呼ぶ. ただし $l_{ij}(x, y)$ はエッジ (i, j) を描画した画像, τ_{ij} はエッジがそれ以外の頂点といくつ画像上で重なっているかを表す交差ペナルティである.

---

* Solving Traveling Salesman Problem with Reinforcement Learning
† Daisuke Yamamoto, Syoma Miki, Graduate School of Kansai University, Suita, Osaka, Japan
‡ Kansai University, Suita, Osaka, Japan

$$v_{ij} = \frac{1}{1+\tau_{ij}} \cdot \frac{\sum_{x=1}^{S_1}\sum_{y=1}^{S_2} p(x,y)\, l_{ij}(x,y)}{\sum_{x=1}^{S_1}\sum_{y=1}^{S_2} l_{ij}(x,y)} \tag{1}$$

2.3 EV-greedy

TSP において, 従来の貪欲法ではすべてのエッジのうち移動コストが最も小さいものを取り込んでいくことで解を構築するが, 移動コストの代わりに優良エッジ値が最大のエッジを選び解を構築する. これを EV-greedy と呼ぶ.

3 提案手法

優良エッジ分布に対して強化学習を適用するために, CNN を用いて生成した解の評価を行う.EV-greedy において次に選択するエッジを決めることを行動とした場合, 学習開始時には価値の高い行動が全くわからないので, 価値の高い行動を選択するためには環境に対して様々な行動を試す必要がある. しかし, EV-greedy は優良エッジ値が最大のエッジを常に選ぶため, 構築を繰り返したときに解の変動が無く一定の行動しか選ばれない. これを改善するために ε-greedy を採用し, 確率 ε で全てのエッジからランダムなものを選び, $1-\varepsilon$ の確率で最も優良エッジ値が高いエッジを選ぶ. これを EV-ε-greedy と呼ぶ.

CNN が出力した優良エッジ分布に対して EV-ε-greedy を用いて生成した解の経路長を報酬として強化学習を行う. 各問題例ごとに報酬と優良エッジ分布の画素値を記憶し, 新たに生成した解の経路長が報酬の指数移動平均よりも短ければ, 優良エッジ分布のその時生成した解に含まれるエッジの各画素に対して画素値が大きくなるように, 長ければ小さくするように学習する. このときの報酬の指数移動平均 $b(k_n)$ を式 (2) に従って更新し, ネットワークの重み θ の変化量 $\Delta\theta$ と, これを決める勾配 g_θ は式 (4), 式 (3) に従って計算する. ここで γ は平滑化係数, B は重みの更新に用いる訓練データの数, $L(\pi_n|k_n)$ は頂点集合 k_n における経路 π_n の長さを表す.

$$b(k_n) \leftarrow \gamma b(k_n) + (1-\gamma) L(\pi_n|k_n) \tag{2}$$

$$\Delta\theta \propto -g_\theta \tag{3}$$

$$g_\theta = \frac{1}{B}\sum_{n=1}^{B}(1 - \frac{b(k_n)}{L(\pi_n|k_n)})\nabla_\theta \sum_{s=1}^{S_x}\sum_{t=1}^{S_y} p_e(s,t|k_n) l_{\pi_n}(s,t|k_n) \tag{4}$$

また, 行動によって得られた経験をランダムにサンプリングした場合勾配が小さい経験も含まれるため学習が進みづらい. そこで, 勾配の絶対値の大きさによって各勾配に優先度を付け, n ステップごとに優先度に基づいてサンプリングを行い学習し, 学習に使った勾配の優先度を変更することによってより学習を進みやすくする.

4 結論

本論文では，平面 TSP における距離に代わる指標として，畳み込みニューラルネットワークにより最適経路画像を近似した優良エッジ分布と，ここから計算される優良エッジ値を用いて, 強化学習を行う手法を提案した．実験結果は発表時提示する.

謝辞 本研究の一部は, JSPS 科研費 18K11484 と, JSPS 科研費 17K01309，関西大学大学院理工学研究科高度化推進研究費，関西大学先端科学技術推進機構「緊急救命避難支援のための情報通信技術に関する研究開発」研究グループの助成を受けている.

参考文献

[1] 三木　彰馬, 榎原　博之 ”深層学習を用いた巡回セールスマン問題の解法” 情報処理学会　情報処理学会論文誌　Vol.60 No.2 (2019)

[2] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglouet, et al.,” Mastering the game of Go without human knowledge ”, Nature 550, pp.354-359, doi:10.1038/nature24270 (2017)

[3] Tom Schaul, John Quan, Ioannis Antonoglou and David Silver, ”Prioritized Experience Replay” arXiv preprint arXiv:1511.05952 (2015)

[4] Horgan, Dan, et al. Distributed prioritized experience replay. arXiv preprint arXiv:1803.00933 (2018)

[5] Bello, Irwan, et al. Neural combinatorial optimization with reinforcement learning. arXiv preprint arXiv:1611.09940, 2016.

[6] Oriol Vinyals, Meire Fortunato, Navdeep Jaitly, ”Pointer Networks ”, NIPS pp.2692-2700 (2015)