ISSN 2186-5647

2-43

深層学習と強化学習の融合における汎化学習

日大生産工　○山根 悠馬　　日大生産工　山内　ゆかり

1. まえがき

近年，機械学習の一種である深層学習や強化学習が画像認識や音声認識，囲碁やテレビゲームのゲームプレイなど多種多様な分野で高い性能を示し，注目を集めている．一方で，これらを車の自動運転などの実際の問題に応用するには，学習法が十分な汎化能力を有していることが必要とされる．ここでいう汎化能力とは，学習時に用いられる訓練データに過適合することなく，未知のデータに対しても対応できる能力を指す．

飯間らは逐次意思決定問題の1つである2次元平面上の経路探索問題に対して，深層学習との併用により汎化能力を向上させた強化学習法1)を提案し，経路探索問題において，狭い平面から学習が困難な広い平面での汎化性向上を可能にした．他の学習法より学習可能数が多く，目標座標到達時刻は少なくなっていることが報告されている．しかし,従来研究では学習したい問題マップで常に高い精度で報酬と状態遷移確率を推定することは困難である．さらに訓練データセットを生成するうえで留意しなければならない点も多く大量の学習データセットが必要などの課題が残されている．

本研究では，上記の安定性を下げているか大部分を減らしリスク回避的なεの値を大きくしたSARSA 2)を用いた強化学習手法を用いることで改良し汎用性を維持しながらも学習正確化と安定化を目的とした手法を提案する．評価には，提案手法と従来研究の学習終了時刻と平均学習収束時間をそれぞれ比較し，学習性能と学習時間の違いについて報告する．

2. 従来研究

従来研究では，深層学習法のネットワークの構造に Value iteration networks 3)を用いる学習法をもとに作られている．そのため，価値関数を深層ネットワークで近似する深層強化学習法ではなく，深層学習と強化学習を逐次的に用いる方法となっている．

従来研究では，強化学習に必要な報酬と状態遷移確率を状態sと行動aを入力とし，深層学習により汎化的に学習する．学習にはミニバッチを適用し，ネットワークの学習パラメータの更新方法には Adam 4)を適用し早く収束する方法をとっている．ネットワークの構造には順伝播の畳み込みニューラルネットワークを用いる．

つぎに，学習した報酬と状態遷移確率の推定値を用いて上記の深層学習時に用いた訓練データセットとは別の訓練データセットにより強化学習を用いて行動価値を学習する．訓練データセットを分ける理由は，同じ訓練データセットを用いる場合行動価値の誤差が少なくなり偏った行動価値ができてしまい未知の行動価値に対する最適方策の予測が誤りやすくなると考えられるためである．この強化学習には,価値反復法を用いるものとし，次式(1)で更新を繰り返し行う．

$$Q(s,a) \leftarrow \mathrm{R(s,a)} + \gamma \mathrm{V(s')}$$
$$\mathrm{V(s)} \leftarrow \max_{\mathrm{a}} \mathrm{Q(s,a)} \qquad (1)$$

価値反復法は行動価値をより正しく学習できるようにするために，価値反復は経路を求めたい問題のデータが与えられた後に実行する．このデータにおける目標座標や障害物座標を利用して学習を行なうものとする．

しかし, 報酬と状態遷移確率を100%の正答率で推定することは困難であり，この結果強化学習により得られる行動価値も誤差を含むことが考えられる．したがって，通常の方法すなわち$\mathrm{arg}max_a Q(s,a)$となる行動で方策を定めると，誤った経路が得られる可能性がある．そこで深層学習を再度用いて，誤差を含む行動価値から最適な行動を教師としたニューラルネットワークで最適方策を学習させる．上記の従来手法のフローチャートをFigure1にて示す．

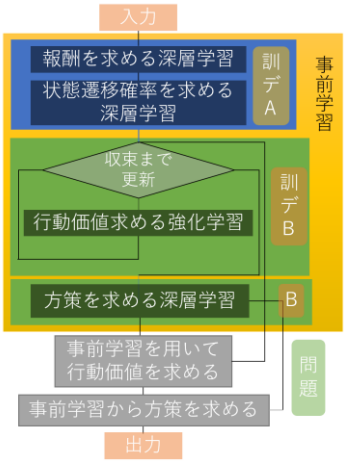

Figure 1　従来手法フローチャート

Generalized Learning by Integrating Deep Learning and Reinforcement Learning

Yuma YAMANE and Yukari YAMAUCHI

3. 提案手法

従来研究では，報酬と状態遷移確率の推定値をもとめる際に訓練データセットを深層学習に入力し推定値を求めている．また価値反復法を行う際，新たなデータセットを用いて学習を行っている．深層学習を行うため学習に多くの良質なデータを必要とし，学習の方向性を絞るために生成するデータの留意点も多い．また，コンピュータリソースも高くなり，時間がかかってしまうという課題がある．

提案手法では，SARSAにより行動価値を求め従来研究の条件を満たす．安定しない状態遷移確率の部分はSARSAで学習することで従来研究より学習が安定化し深層学習で必要なデータセットの削減も図る．

行動価値を求める際，Q学習が多く使われているが，Q学習では収束が早いものの局所解に陥りやすく未知データでの汎用性に欠ける可能性があるためSARSAを本研究では採用する．

SARSAは次式(2)で更新を繰り返し行う．

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r' + \gamma Q(s',a') - Q(s,a)] \quad (2)$$

SARSAではリスク回避的でマップ全体の学習が得られるように状態からのとる行動の行動選択にはε-greedy法に従い，εは大きめの設定としランダム性の高いSARSA を使用する．

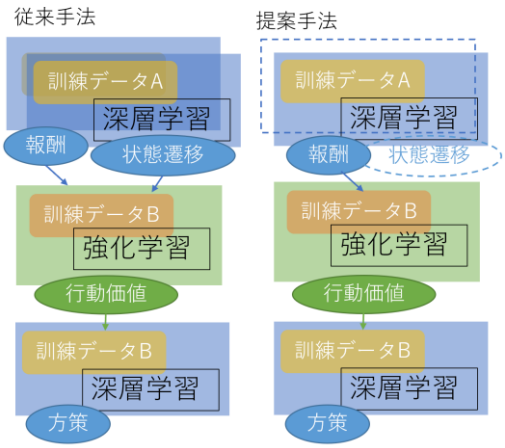

Figure 2　従来手法と提案手法の相違点

4. 実験手法

本実験では，複数の問題データマップの学習を行う．問題マップのマップサイズ$8 \times 8, 28 \times 28$の2次元の経路問題であり，エージェントの選択できる行動の集合AはA＝{左,右,下,上}と設定する．提案手法の汎用性が維持されているか検証を含むため，問題データマップはランダム生成を行う．訓練データセットも同様のマップサイズでランダム生成により作成する，訓練データセットAを900個，訓練データセットBを100個作成する．問題データセットのうちランダムで1つをディスプレイに可視化し最短行動数，学習時間をそれぞれ表示しデータを収集する．強化学習，従来手法，提案手法は分けて動作させデータを収集する．各手法の平均学習時間，学習可能データ数,最短行動数平均を比較し，学習安定性と学習時間,汎用性の違いについて検証する．

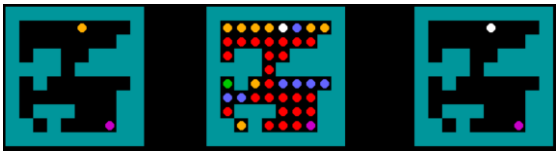

Figure3　強化学習による方策のみ on

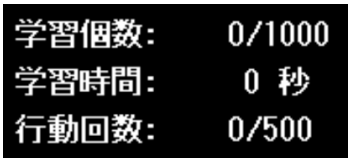

Figure 4　評価データの可視化

5. まとめ

本研究では，学習の汎用性は維持しつつ学習の安定化と高速化を目指した手法を提案した．しかし，他にも課題として本実験では2次元かつエージェントの移動が上下左右の単純な問題データを扱っているが，現実問題では3次元以上での問題が多くエージェントの移動もより複雑な問題であることも多い．

このような現実的な問題にも対応できるよう提案手法を拡張し有用性を検証することが今後の課題としてあげられる．

参考文献

1) 飯間　等，大西　鴻哉，経路探索問題に対して深層学習との併用により汎化能力を向上させた強化学習法，計測自動制御学会論文集，Vol.56, No.10, 455/462（2020）
2) Richard S.Sutton,Andrew G.Barto,三上貞芳,皆川雅章,「強化学習」森北出版（2000）pp.155,171,227
3) Aviv Tamar, Yi Wu, Garrett Thomas, Sergey Levine, and Pieter Abbeel, "Value Iteration Networks" Neural Information Processing Systems (NIPS 2016)
4) D.P. Kingma and J. Ba " Adam: A method for stochastic optimization" Proc. International Conf. on Learning Representations (2015)