6P-03

分布シフトの影響を緩和する深層強化学習におけるモデル平均化手法

A Weight Averaging Method Mitigates the Effect of Distribution Shift in Deep Reinforcement Learning

高橋 快成 *1
Kaisei Takahashi

長沼 大樹 *2*3
Hiroki Naganuma

*1北陸先端科学技術大学院大学
JAIST: Japan Advanced Institute of Science and Technology

*2モントリオール大学
Université de Montréal

*3Mila
Mila - Quebec AI Institute

実用的な深層強化学習のスキームであるロボット制御では、シミュレーション環境において学習し、実際のロボットでの動作を想定する。シミュレーションと実社会での動作における環境差、データの分布シフトによって獲得したエージェントが汎化しないことが喫緊の課題である。近年、アンサンブル手法がこの課題に対して有効であることが示されているが、膨大な計算コストを必要とする。我々は、計算コストの削減と学習の安定化が報告されているアンサンブル手法の近似手法に着目し、この課題に取り組む。Super Mario Bros の異なるステージを実験環境として用いて、疑似アンサンブル手法が深層強化学習における分布シフトの堅牢性へ与える影響について検証を行なった。

1. はじめに

深層強化学習は、ロボティクスの制御分野において数多くの成功を収めている。[Levine 18, Andrychowicz 20] 深層強化学習による制御ポリシーの獲得には、シミュレーション環境で学習を行う Sim-to-Real という手法を用いることでサンプルの収集効率や膨大な学習コストを改善する。しかし、実環境とシミュレーション環境間に生じる環境差や分布シフトによる影響から堅牢な制御ポリシーの獲得が困難であることが知られている。[Zhao 20] この問題への対処法としてアンサンブル手法をベースとする学習手法の有効性が示されているが、膨大な計算コストが必要となる [Lee 21]。

そこで、ニューラルネットワークの重みを平均化するアンサンブル手法の近似手法の有用性が示されている。[Nikishin] 本研究では、深層強化学習におけるアンサンブル手法の近似手法が分布シフトに対する堅牢性に対して与える影響について調査する。

2. 関連研究

Sim-to-Real における環境差や分布シフトへの対応として、学習過程におけるシミュレータ上のパラメータをランダム化する手法 [Tobin 17] や敵対的生成ネットワークやデータ拡張などを活用することで異なる環境へ適応させる手法 [Murooka 21] が提案されている。前者は、事前に決められた特徴に対して過適合を防ぐような役割があるが、実世界におけるシミュレータ上で想定されていない要素に対応ができずに失敗してしまうという問題点がある。後者は、ドメイン適応の問題設定であり、適応先の教師なしデータに学習へアクセスすることが求められ、本研究では適応先のデータにアクセスできないドメイン汎化の設定をスコープとするため用いることができない。分布シフト下においてドメイン汎化として定式化可能な問題設定での堅牢性を高めるため、次節に示すアンサンブル手法が注目されている。

2.1 アンサンブル学習

アンサンブル手法は深層強化学習の安定性と性能向上に貢献することが報告されている [Kurutach 18, Yu 20]。アンサンブル手法に基づいたベルマンバックアップの重み付手法によって既存のオフポリシー強化学習アルゴリズムの性能を改善させることに成功した [Lee 21]。また、アンサンブル手法を深層強化学習において環境変化が起こるような場合であっても、機能することが報告されており [Sogabe 22]、分布シフトに対して堅牢なエージェントの獲得が期待できる。

2.2 擬似アンサンブル学習

確率的重み付け平均法 (SWA: Stochastic Weight Averaging) [Izmailov 18] は、アンサンブル学習の近似手法であり、モデルの出力をアンサンブルするのではなく、モデルのパラメータ自体の平均化を行う。SWA は、教師あり学習における画像の分類問題に対して、損失関数の形状に着目したアンサンブル手法 FGE [Garipov 18] の近似手法として開発され、画像分類 [Izmailov 18]・言語 [Maddox 19] にも活用されている。

1. 章で示した通り、深層強化学習へのアンサンブル学習の適応は、学習の安定化に寄与する一方で、アンサンブル手法は通常の手法よりも多くの計算資源が必要となる問題が指摘されている [Lee 21]。この問題の解決のため、Evgenii らは、SWA を 強化学習における Advantage Actor-Critic (A2C) [Barto 83] の問題設定に適応し、学習の不安定性を解消したことを報告している [Nikishin]。

$$W_{swa} = \frac{W_{\text{swa}} \cdot n_{\text{models}} + W}{n_{\text{models}} + 1}$$

3. 実験

機械学習フレームワークとしては PyTorch*1 を、ベンチマークとしては Super Mario Bros*2 を用いた。DNN モデルは CNN を用いて学習を行った。エージェントの学習手法としては、Super Mario Bros に広く用いられている Double Deep Q Network [Van Hasselt 16] を用いる。最適化手法は Adam を用いており、SWA の平均化周期は $c = 100$ とする。本実験では、stage1-1 で学習を行った同程度の水準を示す SWA ありとなしの 2 つのモデルを作成し、分布シフトの環境として Stage1-2, 1-3, 1-4 で評価を行った。エージェントは 7000 と 13000 episode の学習を行い、学習終了後にそれぞれの Stage で獲得できた報酬を表 1a1b に示す。比較するモデルは Stage1-1 において SAW の方が低い報酬結果となっている。しかしながら、評価を行う Stage1-2, 1-3, 1-4 において SWA は Adam と同等かそれを上回る獲得報酬となった。

連絡先: s2250003@jaist.ac.jp

*1 https://pytorch.org/
*2 https://github.com/Kautenja/gym-super-mario-bros

(a) 7000episode

Stage	Adam	SWA
1-1	**1144**	627
1-2	128	**134**
1-3	115	**234**
1-4	169	169

(b) 13000episode

Stage	Adam	SWA
1-1	**630**	610
1-2	137	137
1-3	248	**374**
1-4	169	169

表 1: 学習終了後における各 Stage ごとの獲得報酬

4. 考察

表 1a の結果から、Stage1-2 および Stage1-3 において SWA が優位であった要因として、1-1 で獲得した方策が Adam で獲得した方策に比べ、堅牢性の高いものであったと考えられる。また、SWA が 1-4 において Adam に比べ優位でなかった要因として、Stage1-4 では、新たな種類のトラップなどが出現し、1-1 で獲得した方策では対応することが困難であったと推察できる。また、表 1a, 表 1b の結果から、通常の Adam の学習では、学習ドメインにおいて高い獲得報酬であっても、分布シフト先の環境ではうまく汎化するとは限らない結果が示唆された。

5. おわりに

本研究は、深層強化学習において、異なる未知環境に対して堅牢なエージェント獲得を目的として、計算コスト・メモリ使用量などの側面で実応用の可能性が期待されるアンサンブル学習の近似手法に着目し、分布シフトに対する堅牢性に対する影響について調査した。実験の結果、アンサンブル学習の近似手法である SWA が近しい特性を持つ環境における分布シフトに対して堅牢なエージェントの獲得につながることを示唆する結果を得た。

今後の課題として、より環境差の大きい分布シフトに対して堅牢なエージェントを獲得するためのアルゴリズムの開発が必要である。また、適応先のデータにアクセスできるようなドメイン適応の問題設定においても比較評価が必要である。

参考文献

[Andrychowicz 20] Andrychowicz, O. M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al.: Learning dexterous in-hand manipulation, *The International Journal of Robotics Research*, Vol. 39, No. 1, pp. 3–20 (2020)

[Barto 83] Barto, A. G., Sutton, R. S., and Anderson, C. W.: Neuronlike adaptive elements that can solve difficult learning control problems, *IEEE transactions on systems, man, and cybernetics*, No. 5, pp. 834–846 (1983)

[Garipov 18] Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G.: Loss surfaces, mode connectivity, and fast ensembling of dnns, *Advances in neural information processing systems*, Vol. 31, (2018)

[Izmailov 18] Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G.: Averaging weights leads to wider optima and better generalization, *arXiv preprint arXiv:1803.05407* (2018)

[Kurutach 18] Kurutach, T., Clavera, I., Duan, Y., Tamar, A., and Abbeel, P.: Model-ensemble trust-region policy optimization, *arXiv preprint arXiv:1802.10592* (2018)

[Lee 21] Lee, K., Laskin, M., Srinivas, A., and Abbeel, P.: Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning, in *International Conference on Machine Learning*, pp. 6131–6141PMLR (2021)

[Levine 18] Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., and Quillen, D.: Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection, *The International journal of robotics research*, Vol. 37, No. 4-5, pp. 421–436 (2018)

[Maddox 19] Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G.: A simple baseline for bayesian uncertainty in deep learning, *Advances in Neural Information Processing Systems*, Vol. 32, (2019)

[Murooka 21] Murooka, T., Hamaya, M., Drigalski, von F., Tanaka, K., and Ijiri, Y.: Exi-net: Explicitly/implicitly conditioned network for multiple environment sim-to-real transfer, in *Conference on Robot Learning*, pp. 1221–1230PMLR (2021)

[Nikishin] Nikishin, E., Izmailov, P., Athiwaratkun, B., Podoprikhin, D., Garipov, T., Shvechikov, P., Vetrov, D., and Wilson, A. G.: Improving stability in deep reinforcement learning with weight averaging

[Sogabe 22] Sogabe, T., Malla, D. B., Chen, C.-C., and Sakamoto, K.: Attention and masking embedded ensemble reinforcement learning for smart energy optimization and risk evaluation under uncertainties, *Journal of Renewable and Sustainable Energy*, Vol. 14, No. 4, p. 045501 (2022)

[Tobin 17] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world, in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30IEEE (2017)

[Van Hasselt 16] Van Hasselt, H., Guez, A., and Silver, D.: Deep reinforcement learning with double q-learning, in *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30 (2016)

[Yu 20] Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T.: Mopo: Model-based offline policy optimization, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 14129–14142 (2020)

[Zhao 20] Zhao, W., Queralta, J. P., and Westerlund, T.: Sim-to-real transfer in deep reinforcement learning for robotics: a survey, in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 737–744IEEE (2020)