F-016

分布シフト下で堅牢な深層強化学習のための モデルのマルチドメイン平均化

高橋 快成[1,a)]　長沼 大樹[2,3,b)]
Kaisei Takahashi[1,a)]　Naganuma Hiroki[2,3,b)]

1. はじめに

深層強化学習を用いて、ロボットに複雑な制御ポリシーを学習させる際には、シミュレーション環境を学習ドメインとして用いる Sim-to-Real のアプローチが広く用いられている [1]。しかし、学習ドメインで獲得した制御ポリシーは、ターゲットドメインである実環境下で適切に機能しないことが課題として挙げられている [12]。これは、ドメイン間における分布シフトの影響が原因であり、課題の解決にはドメインギャップに対して堅牢な学習が必要となる。

この課題に対するアプローチの一つとしてアンサンブル手法を用いた学習がある。アンサンブル手法は、不確実性に対する堅牢性が高く、ドメインギャップへの対策として期待がされている。しかし、アンサンブル手法では複数の弱学習器が必要であることから、膨大な学習コストが必要となる。そこで、注目がするのがニューラルネットワークの重みを平均化するモデル平均化である [8]。しかし、単一の学習ドメインを用いたエージェントに対するモデル平均化では、分布シフトに対して堅牢な学習を実現することは難しい。そこで、本研究では、複数の学習ドメインで学習したエージェントに対してモデル平均化を行うことで、深層強化学習における分布シフトに対する堅牢性を改善することができるのかについて調査する。

2. アンサンブル学習とモデル平均化

2.1 アンサンブル学習

アンサンブル学習は、強化学習に適応することで、学習の安定性と性能向上に貢献する [5, 11]。また、深層強化学習におけるアンサンブル手法の適応は、不確実な環境下において、高い性能を示すことが報告されており [10]、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。しかし、深層強化学習におけるアンサンブル学習の適応は、弱学習器分の学習コストの増加が必要 [6] となり、深層強化学習の持つ膨大な学習コストの問題をさらに深刻化させる。

1 北陸先端科学技術大学院大学/ JAIST
2 モントリオール大学/ Université de Montréal
3 Mila/ Montreal Institute for Learning Algorithms
a) s2250003@jaist.ac.jp
b) naganuma.hiroki@mila.quebec

2.2 モデル平均化

確率的重み付け平均法（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 [4] の近似手法として開発されたアンサンブル学習の近似手法であり、ニューラルネットワークの重みパラメータの平均化を行う。Evgenii らは、深層強化学習アルゴリズムである Advantage Actor-Critic(A2C) [2] と Deep Deterministic Policy Gradient（DDPG）[7] に対して SWA を適応し、学習の不安定性を解消したことを報告している [8]。

2.3 複数の学習ドメインに対するモデル平均化

アンサンブル学習の性能向上には、弱学習器の多様性が重要である [3]。しかし、単一の学習ドメインにおける SWA の適応の場合、平均化するパラメータの多様性は低く、分布シフトに対する堅牢性の改善は期待しにくい。そこで、本研究では新たな試みとして、複数の学習ドメインを定期的に切り替えながら学習を行う単一のエージェントに対して SWA を適応する Multi-Domain SWA（MDSWA）を提案する。提案手法により獲得した複数ドメインにおける多様なパラメータを平均化することで、分布シフトに対する堅牢性の改善につながるかを検証する。

$$W_{\text{MultiDomain-SWA}} = \sum_{d \in D} \sum_{t \in T} W_{dt} \qquad (1)$$

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。ベースの学習となる深層強化学習アルゴリズムは、Super Mario Bros に広く用いられている Proximal Policy Optimization（PPO）[9] を用いる。MDSWA の学習は、9000 エピソード実施し、最適化手法は Adam、SWA の平均化周期は 200 とする。stage1-1、1-2、1-3 の 3 ステージを学習ドメインとし、50 エピソード周期でそれらを切り替えながら学習したエージェントに 3000 エピソード終了時点から SWA を適応する。本実験では、マルチドメインに対して Adam の

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

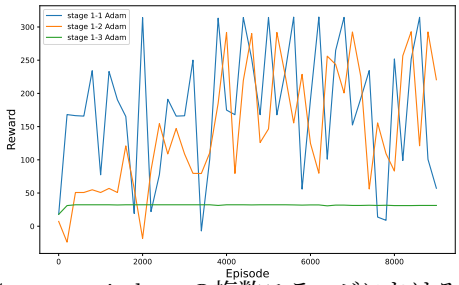

図 1: Super mario bros の複数ステージにおける Adam の獲得報酬の推移（200 エピソードごと）

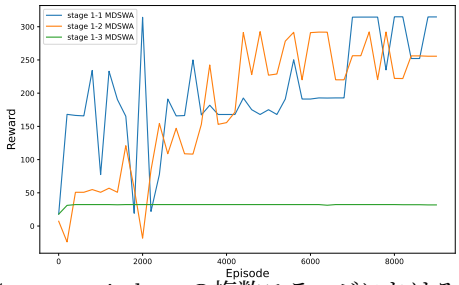

図 2: Super mario bros の複数ステージにおける MDSWA の獲得報酬の推移（200 エピソードごと）

みで学習したモデルと SWA を適応した MDSWA の 2 つのモデル作成し、Stage1-1、1-2、1-3、2-1 で評価を行った。評価実験の結果を表 1 に、学習ドメインである Stage1-1、1-2、1-3 における獲得報酬の推移を図 1、2 に示す。

まず、未知のドメインについて述べる。表 1 の通り、未知のドメインである Stage2-1 では MDSWA が Adam より多くの報酬を得た。しかし、Adam、MDSWA いずれも低い獲得報酬であり、Stage の序盤で失敗している。

次に、学習ドメインにおける獲得報酬について述べる。表 1 より、MDSWA は Stage1-1、1-2 の複数ステージで高い水準の獲得報酬を示しており、1-1 についてはクリア、1-2 についてはクリアの目前まで学習を実現することができた。Stage1-3 については、同等の報酬に留まり、いずれも低い獲得報酬である。また、図 1、2 より Adam のみのマルチドメイン学習は、Stage1-1 の獲得報酬が高い時 Stage1-2 の獲得報酬が低くなるトレードオフ現象を複数のエピソードで観測した。しかし、SWA を用いた場合この現象は緩和され、Stage1-1 と 1-2 ともに高い獲得報酬で推移する。

Stage	1-1	1-2	1-3	2-1
Adam	57.1	220.6	31.3	20.8
MDSWA	314.7**(C)**	255.6	31.8	34.9

表 1: 学習終了後における各 stage における獲得報酬。**(C) はステージのクリア表す。**

4. おわりに

MDSWA は、複数ドメインにおける分布シフトに対する堅牢性を改善することを示唆する結果を得ることができた。特に、学習ドメイン間における分布シフトの影響を緩和することが可能であり、複数のドメインに対する性能を向上させる。一方で、未知の分布シフトに対する堅牢性の改善には、課題が存在する。この原因として、未知の分布シフト環境に存在する異なる特性を持った構造やギミックに対応できる多様なモデルパラメータを獲得できなかったことが挙げられる。

今後の課題として、未知の分布シフトに対して堅牢なエージェントの学習を実現する手法の研究が必要である。

謝辞 本研究は、令和 5 年度 東京工業大学学術国際情報センターの「TSUBAME 若手・女性利用者支援制度」および学際大規模情報基盤共同利用・共同研究拠点の「萌芽型共同研究 (EX23402)」の支援を受けた。

参考文献

[1] 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).

[2] 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).

[3] Dietterich, T. G.: Ensemble methods in machine learning, *International workshop on multiple classifier systems*, Springer, pp. 1–15 (2000).

[4] 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).

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

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

[7] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. and Wierstra, D.: Continuous control with deep reinforcement learning, *arXiv preprint arXiv:1509.02971* (2015).

[8] 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.

[9] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O.: Proximal policy optimization algorithms, *arXiv preprint arXiv:1707.06347* (2017).

[10] 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).

[11] 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).

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