F-042

深層強化学習における疑似アンサンブル手法の探索と活用のトレードオフ

高橋 快成 ♯,a)　長沼 大樹 ♯2,3,b)
Kaisei Takahashi ♯,a)　Naganuma Hiroki ♯2,3,b)

1. はじめに

深層強化学習エージェントは、近年、ゲームを始めロボットの制御タスクにおいて多くの成功を収めている [1, 15]。適切な方針を学習するためには環境との多くの相互作用が必要となり、ロボットアームの制御などのシナリオでは、実環境での学習には膨大な時間を要するため、シミュレーション環境での学習が行われる [1]。これらの環境差への対処法として、アンサンブル手法をベースとする学習手法はロバストなエージェントの獲得につながるが、膨大な計算資源を要する [11]。

そこで、アンサンブル手法の近似手法として、重み自体を平均化する手法の有効性が示されている [14]。この際用いられる平均化間隔 c はヒューリステックに決定されている。本研究では、c が探索と活用のトレードオフを制御することに着目し、学習に及ぼす影響を調査する。

2. 探索と活用のトレードオフ

深層強化学習においては、学習過程における適切でない行動選択の影響によって高い報酬を得た行動を忘却してしまい、学習が不安定となる問題が知られている [14]。強化学習で最適な行動を学習するためには、最適な行動を探る「探索」と既知の行動の中で最適な行動をとる「活用」のバランスを取りながら意思決定を行っていく必要がある。しかし、両者の間にはトレードオフの関係が成立している。「探索」を優先しすぎた場合、最適な行動以外の行動を多く選択する可能性が上がり、学習が不安定になる。一方で、活用を優先しすぎた場合、十分な探索が行われないことから、最適でない行動に収束してしまうリスクが高まる。ことから、これらのバランスは強化学習における重要な課題として知られている [9]。

3. アンサンブル学習

学習安定化に限らず性能向上のため、深層強化学習においても、アンサンブル手法が用いられる [10, 16]。アンサンブル学習の性能向上には、弱学習機の多様性が必要であることが知られている [5]。この多様性の獲得は、様々な手段で実現することができ、データのランダムな部分集合による学習 [3]、ランダムな特徴量の部分集合による学習 [4]、ブースティング的なアプローチ としては学習データの重み付けを操作する手法 [6] などが挙げられる。ニューラルネットに限定した場合、同一のアーキテクチャを異なる初期化およびミニバッチの順序で複数回学習させることで多様性が得られることが、確率的勾配降下法のランダム性とニューラルネットワークの損失関数が非凸であるという側面から示されている [12]。

4. 深層強化学習と確率的重み平均化法

確率的重み付け平均法 (SWA: Stochastic Weight Averaging) [8] 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 を用いた学習手法をアルゴリズム 1 に示す。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 [7] の近似手法として開発され、画像分類 [8]・

Algorithm 1: Stochastic Weight Averaging for RL

Parameters : Model parameter $\theta \in \mathbb{R}^p$
Require : η: Learning rate, t: Steps, $\mathcal{L}(\theta)$: loss function, c: Cyclic period for SWA
Ensure : Initialize θ as θ_0 ,and θ_{swa} as θ_0
while *end of training* **do**
  `// Get learning rate` η `w.r.t` t
  $\eta \leftarrow$ LR-Scheduler(t) ;
  Forward and Backward ;
  `// Update model parameter` θ_t `by using gradient`
  $\theta_{t+1} \leftarrow \theta_t - \eta L(\theta_t)$;
  if $t \bmod c = 0$ **then**
    `// Update` θ_{swa} `once every` c `training steps`
    $N_{\text{models}} \leftarrow i/c$;
    $\theta_{\text{swa}} \leftarrow \frac{\theta_{\text{swa}} \times N_{\text{models}} + \theta_{t+1}}{N_{\text{models}} + 1}$;
  end
end
Return : θ_{swa}

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
♯ denotes equal contribution

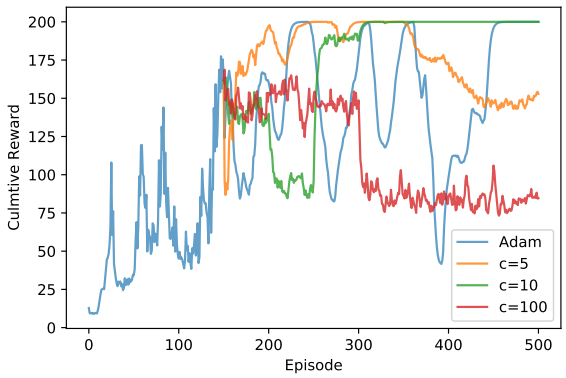

図 1: Average cumulative rewards of A2C for CartPole

言語 [13] 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 [11]。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) [2] 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 [14]。ここで、モデルの平均化周期 c は、疑似的なアンサンブルを行う弱学習器の数 (探索) と、それぞれの弱学習器自体の多様性 (活用) を制御していると解釈できる。本研究では、c が実験的にもトレードオフを制御するのかを検証し、学習に及ぼす影響を調査する。

5. 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 Cart-Pole[*2]を用いた。DNN モデルは三層の MLP 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、Evgenii ら [14] に倣い A2C を用いる。最適化手法は Adam を用いており、学習全体の 30%から SWA を適用した。図 1 に、平均化周期 c を変化させた場合の報酬に関する比較実験結果を示す。Adam 単体を使った場合、学習が安定しないのに対し、$c = 10$ の場合の SWA の学習は安定する結果となった。$c = 5$ の場合は、平均化モデルが多いこともあり学習が不安定化し、$c = 100$ の場合は最適な行動に収束しなかった。

6. おわりに

本研究は、深層強化学習において、安定的な学習・ロバスト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、平均化周期 c が学習にもたらす影響を実験的に調査した。実験の結果、平均化周期 c が探索と活用のトレードオフを制御することを示唆する結果を得た。

今後の課題として、報酬推定を行う逆強化学習 [17] を適応することでトレードオフのコントロールを適応的に行う手法の開発や、SWA を適応したエージェントの、環境シフト下での性能評価が必要である。

*1 https://pytorch.org/
*2 https://github.com/openai/gym/wiki/CartPole-v0

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] Breiman, L.: Bagging predictors, *Machine learning*, Vol. 24, No. 2, pp. 123–140 (1996).

[4] Breiman, L.: Random forests, *Machine learning*, Vol. 45, No. 1, pp. 5–32 (2001).

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

[6] Freund, Y.: Boosting a weak learning algorithm by majority, *Information and computation*, Vol. 121, No. 2, pp. 256–285 (1995).

[7] 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).

[8] 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).

[9] Kearns, M. and Singh, S.: Near-optimal reinforcement learning in polynomial time, *Machine learning*, Vol. 49, No. 2, pp. 209–232 (2002).

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

[11] 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).

[12] Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D. and Batra, D.: Why M heads are better than one: Training a diverse ensemble of deep networks, *arXiv preprint arXiv:1511.06314* (2015).

[13] 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).

[14] 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.

[15] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al.: Mastering the game of Go with deep neural networks and tree search, *nature*, Vol. 529, No. 7587, pp. 484–489 (2016).

[16] 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).

[17] Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K. et al.: Maximum entropy inverse reinforcement learning., *Aaai*, Vol. 8, Chicago, IL, USA, pp. 1433–1438 (2008).