Atari ゲームに対する Transformer ベース強化学習のロバスト性検証

Robustness Verification of Transformer-based Reinforcement Learning for Atari Games

髙野 剛志 [*1] 計良 宥志 [*2] 川本 一彦 [*2]
Tsuyoshi Takano Hiroshi Kera Kazuhiko Kawamoto

[*1]千葉大学大学院融合理工学府
Graduate School of Science and Engineering, Chiba University

[*2]千葉大学大学院工学研究院
Graduate School of Engineering, Chiba University

In this study, we examine the robustness of Transformer-based offline reinforcement learning. We train the state data for offline reinforcement learning with noise and evaluate the performance in training under noise. In the evaluation experiments, we compare the performance of four different Atari games (Breakout, Pong, Qbert, and Seaquest) in terms of scores on five evaluation tests (Clean, Gaussian, Shot, Impulse, and Speckle). The experimental results showed that the Atari game scores were lower on all noise evaluation tests for normal training (clean). The results showed that the Atari game score tended to improve when data augmentation training with a noise system was introduced during training. This result indicates the vulnerability of Atari games to noise evaluation tests and the robustness of Atari games improved by data augmentation training.

1. はじめに

Transformer[1] は，ChatGPT を代表とする言語タスクにおいて大きな成果を出している．この Transformer を画像タスク [2] やロボットタスク [3] に応用した研究があり，応用先は多岐にわたることが知られている．その中でも Transformer を強化学習に応用したもの [4] があり，言語タスクと同様に系列データ (軌道) を扱うことができる．Transformer を応用した強化学習は従来手法と比較して高い性能を示す一方で，脆弱性やロバスト性は我々の知る限りでは明らかにされておらず，検証の必要性がある．

Atari ゲームにおける深層強化学習には，敵対的摂動に対する脆弱性が存在する [5]．例えば、敵対的摂動によってエージェントを目標状態に誘導することができたり [6]，敵対的摂動とランダムノイズの両方で方策の性能低下を引き起こすことができる [7]．このような摂動に対してロバスト性を高めることは重要な課題である．敵対的摂動は意図したノイズを付与するが，普遍的破損 [8] のような各種ノイズに関しては，意図せずノイズが付与される可能性があり，検証の必要性がある．

本研究では，Atari ゲームにおいて各種ノイズに対する脆弱性の検証とデータ拡張訓練によるロバスト性向上の検証を目的とする．検証実験における通常訓練とデータ拡張訓練の比較により，データ拡張訓練によるロバスト性向上が見られたことを示す．

本研究の主な貢献は以下の通りである．

- 普遍的破損における各種ノイズ評価テストにて脆弱性が存在することを示した．
- データ拡張訓練により，ロバスト性が向上することを示した．

連絡先: 川本 一彦，kawa@faculty.chiba-u.jp

2. 関連研究

2.1 強化学習におけるロバスト性

強化学習におけるロバスト性向上を提案した研究がある．Huan らは，状態空間に敵対的摂動を加える敵対者と，強化学習 (PPO, DDPG, DQN) を交互に訓練することで，ロバスト性を向上させる手法を提案している [9]．また Tuomas らは，状態空間への敵対的摂動に対して，敵対的損失を利用することで強化学習 (DQN,A3C,PPO) のロバスト性を向上させる手法を提案した [10]．これらの研究は，Atari ゲームにおける実験と敵対的摂動に対するロバスト性向上の提案という点で共通点がある．その一方で，Trasformer ベース強化学習における検証はされておらず研究の余地がある．

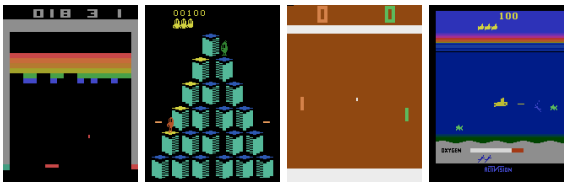
2.2 強化学習におけるデータ拡張

強化学習におけるデータ拡張とは，画像ベースおよび状態ベースの入力に対してデータ拡張を適用し，強化学習アルゴリズムを変えることなく性能を向上させる手法である [11]．

強化学習におけるデータ拡張の有効性を示す研究がある．CURL[12] ではサンプル効率向上のため，強化学習の文脈において対照学習を用いたデータ拡張を提案した．また，強化学習における観測画像に対して，回転や反転などの一般的なデータ拡張手法を適用することで，モデルの汎化能力が向上することが報告されている [13]．画像分野ではデータ拡張手法として普遍的破損 [8] によるロバスト性検証が行われている一方で，強化学習に対して普遍的破損を適用したロバスト性は検証されていない．

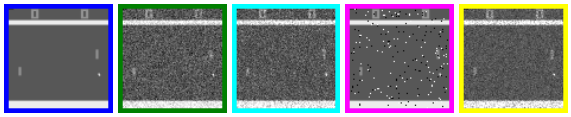
3. データ拡張を用いた訓練

本節では，Atari ゲームの訓練と評価テストに使用する 4 種類のノイズ摂動と訓練方法を述べる．実験環境として，図 1 のゲーム環境 Atari の中から 4 種類のゲームタスクを使用する．Atari ゲームの訓練には，Transformer ベース強化学習の 1 つである Decision Transformer (DT)[4] を用いる．また，ノイズ手法は普遍的破



(a) Breakout (b) Qbert (c) Pong (d) Seaquest

図 1: 使用する 4 種類のゲームタスク



(a) Clean (b) Gaus (c) Shot (d) Imp (e) Spe

図 2: Pong の訓練で使用する観測画像の例

損 [8] に基づいて実装している．本節で説明するノイズ手法は，訓練データと評価テストに使われる．訓練データでの適用例を図 2 に示し，評価テストでの適用例を図 3 に示す．

方策は，ゲーム攻略が可能になるように合計約 20000 回 (iteration) の訓練により学習される．1 epoch の学習に対して約 4000 回の学習を 5 epoch 行うことで約 20000 回の訓練となる．

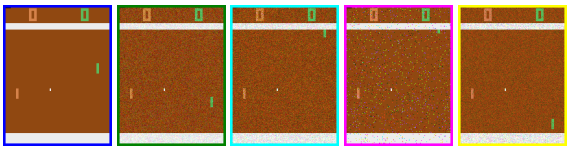
訓練は，図 2 の各種ノイズで DT モデルを訓練する．訓練には，DQN-Replay データセット [14] を使用する．このデータセットの観測データに対して，各種ノイズを適用したデータセットをそれぞれ Gaussian, Shot, Impulse, Speckle として使用する．DQN-Replay は 100 万 ×50 の 5000 万タプル $(a_t,\ s_t,\ r_t,\ s_t+1)$ のオフラインデータセットであり，全てを使用するとメモリオーバーする．そこで，DQN-Replay 1% データセットとしてリプレイバッファ容量を 50 万タプルとして訓練する．リプレイバッファに貯められるデータは 5000 万タプルの中からランダムに選択される．

4. データ拡張訓練によるロバスト性検証

本節では，前節で述べた方法に従い，データ拡張訓練によるロバスト性を検証する．各訓練による方策の評価テストには図 3 の Clean，Gaussian，Shot，Impulse，Speckle の 5 種類を用いる．1 epoch の学習終了後，10 回の評価テストを行う．それを 5 epoch 分 (計 50 回) 試行した際のスコアをとり，3 つの seed 値によって算出される平均スコアと標準偏差から表を作成する．表の太字は，各評価テストにおいて一番スコアが高い数値を示す．また，3 つの seed 値による結果から訓練と評価テストの組み合わせにおける箱ひげ図を作成する．横軸は訓練方法を表し，縦軸はスコアを表す．箱ひげ図の色によって各種ノイズにおける評価テストを表している．箱ひげ図は白丸で平均値を示し，外れ値は除外している．

4.1 脆弱性評価

ここでは，通常訓練 (Clean) に対して各種ノイズにおける評価テストを適用した場合の性能を比較する．



(a) Clean (b) Gaus (c) Shot (d) Imp (e) Spe

図 3: Pong の Clean 訓練に対する各テストの例

表 1, 2, 3, 4 に示すように，4 種類のゲームタスクにおける Clean 訓練に対する各種ノイズの評価テストでは，Gaussian, Shot, Impulse, Speckle の全てでスコア低下が確認された．

また，図 4, 5, 6, 7 に示す箱ひげ図を見ると，評価テストが Clean の場合と比較してスコア分布が大幅に低下する場合があることが確認された．特に図 6 に示す Pong タスクの場合におけるスコア分布の大幅な低下が顕著である．

この結果から，通常訓練に対する評価テストとして各種ノイズを適用すると脆弱性を示すことが明らかとなった．

4.2 ロバスト性評価

ここでは，各評価テストに対して各訓練手法を適用した際のスコアを評価する．

表 1 の Breakout では Clean, Gaussian, Shot 評価テストにおいて，対応する訓練手法で最高スコアを示すことが確認された．また，表 4 の Seaquest では Gaussian, Speckle 評価テストにおいて，対応する訓練手法で最高スコアを示すことが確認された．表 2 の Qbert では，各評価テスト全てにおいて Impulse 訓練が最高スコアを示すことが確認された．これらの結果から Breakout, Qbert, Seaquest では各評価テストに対応する訓練手法が最善であるとは限らないことが言える．その一方で，「各評価テストに対する Clean 訓練でのスコア」と「各評価テストに対応する訓練手法でのスコア」を比較すると，後者においてスコア向上の傾向が確認された．

表 3 の Pong では，各評価テストに対応する訓練手法で最高スコアを示すことが確認された．また，図 6 に示す Pong の箱ひげ図を見ると，Clean 訓練に対する各評価テストのノイズ系で，スコア分布が低下している．そこに各評価テストに対応するデータ拡張訓練を適用することでスコア分布が向上していることがわかる．

この結果からゲームタスクに依存するものの，各評価テストに対してデータ拡張訓練を適用することでロバスト性が向上することが示された．

5. おわりに

評価テストに各種ノイズを加えることで脆弱性を示し，その対策としてデータ拡張訓練を導入し，ロバスト性を検証した．今後は，データ拡張訓練と評価テストに敵対的摂動を加えて調査を進める．

謝辞

本研究は JSPS 科研費 JP22H03658 および電気通信普及財団の助成を受けたものです．

表 1: Breakout における 3seed 値の平均と標準偏差

	Test				
Train	**Clean**	**Gaussian**	**Shot**	**Impulse**	**Speckle**
Clean	**48.00 ± 41.65**	2.486 ± 3.199	38.98 ± 34.03	1.72 ± 2.47	38.24 ± 28.01
Gaussian	12.10 ± 9.942	**19.18 ± 14.94**	13.28 ± 10.03	**2.0 ± 3.2**	13.65 ± 21.21
Shot	47.26 ± 41.62	1.693 ± 2.256	**40.68 ± 29.88**	1.853 ± 2.546	**44.72 ± 38.97**
Impulse	1.666 ± 1.878	2.226 ± 3.175	2.533 ± 2.644	1.786 ± 1.791	2.213 ± 2.504
Speckle	39.95 ± 28.13	2.26 ± 2.99	32.98 ± 23.48	1.786 ± 2.149	34.34 ± 27.01

表 2: Qbert における 3seed 値の平均と標準偏差

	Test				
Train	**Clean**	**Gaussian**	**Shot**	**Impulse**	**Speckle**
Clean	2205.5 ± 2770.9	346.16 ± 212.71	1100.0 ± 1778.3	218.16 ± 153.70	2163.1 ± 2876.4
Gaussian	6303.8 ± 4633.0	1103.0 ± 1746.2	3150.6 ± 3505.0	335.83 ± 225.76	3831.8 ± 3929.2
Shot	3934.3 ± 4230.0	870.66 ± 1368.4	1683.8 ± 2214.0	246.0 ± 169.6	2301.8 ± 3018.6
Impulse	**6660.1 ± 4614.6**	**1553.5 ± 2128.7**	**4079.6 ± 3461.7**	**882.16 ± 1188.8**	**4831.1 ± 4165.1**
Speckle	2602.5 ± 2961.8	472.0 ± 688.3	1745.3 ± 2641.7	266.83 ± 191.43	1539.6 ± 2373.1

表 3: Pong における 3seed 値の平均と標準偏差

	Test				
Train	**Clean**	**Gaussian**	**Shot**	**Impulse**	**Speckle**
Clean	**1.826 ± 17.21**	-20.14 ± 1.035	-19.92 ± 1.233	-20.56 ± 0.7156	-19.99 ± 0.9899
Gaussian	-7.693 ± 17.52	**0.4733 ± 16.07**	-9.033 ± 15.84	-9.813 ± 10.17	1.68 ± 17.22
Shot	1.593 ± 18.02	-1.486 ± 15.68	**-0.5066 ± 16.33**	-12.40 ± 8.369	0.1533 ± 16.96
Impulse	1.0 ± 16.82	-0.3533 ± 15.88	-1.206 ± 15.38	**-1.773 ± 14.84**	0.98 ± 17.04
Speckle	1.746 ± 18.24	-16.86 ± 4.016	-12.16 ± 8.311	-18.32 ± 2.412	**2.34 ± 16.68**

表 4: Seaquest における 3seed 値の平均と標準偏差

	Test				
Train	**Clean**	**Gaussian**	**Shot**	**Impulse**	**Speckle**
Clean	845.7 ± 424.0	502.1 ± 246.9	678.4 ± 335.8	394.6 ± 173.5	814.6 ± 409.3
Gaussian	701.8 ± 353.1	**719.3 ± 316.5**	715.2 ± 319.8	**664.8 ± 277.4**	686.8 ± 309.5
Shot	800.6 ± 372.0	706.9 ± 344.7	692.4 ± 373.4	535.3 ± 217.5	675.6 ± 379.6
Impulse	577.0 ± 305.7	505.0 ± 273.2	581.3 ± 319.0	549.8 ± 287.3	563.0 ± 300.4
Speckle	**883.7 ± 430.6**	680.2 ± 308.0	**835.0 ± 388.0**	435.3 ± 209.3	**814.9 ± 391.1**

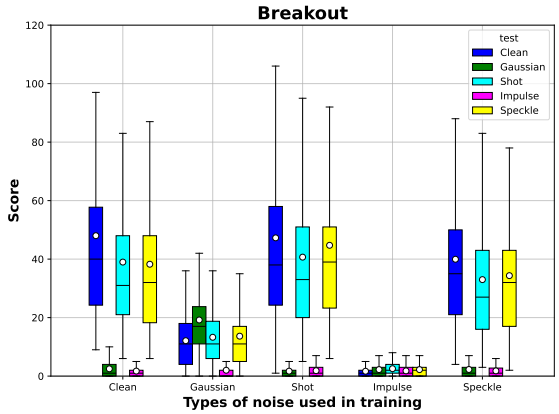


図 4: Breakout での各ノイズ訓練に対するスコア

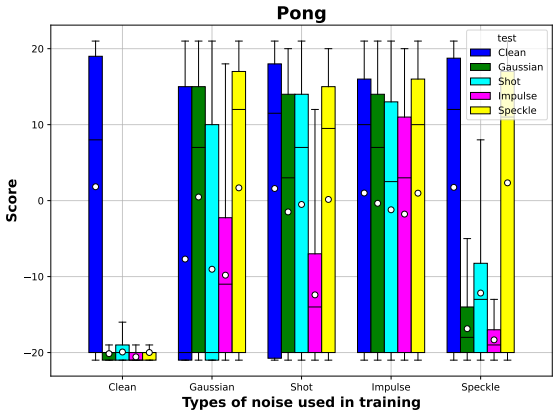


図 6: Pong での各ノイズ訓練に対するスコア

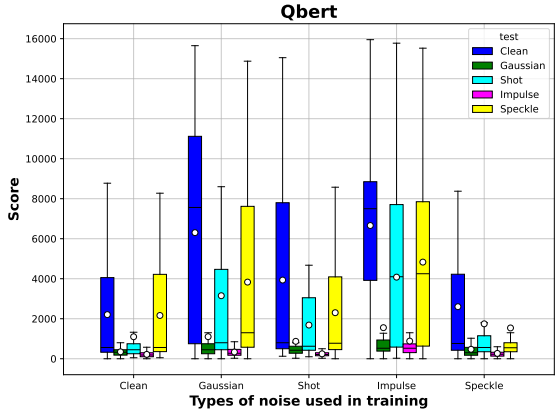


図 5: Qbert での各ノイズ訓練に対するスコア

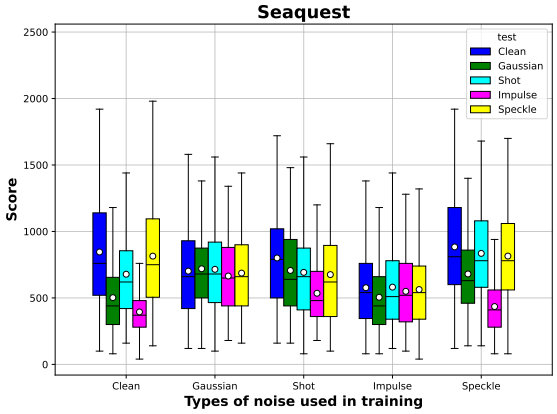


図 7: Seaquest での各ノイズ訓練に対するスコア

参考文献

[1] Ashish Vaswani et al. Attention is all you need. *Advances in neural information processing systems*, 30:5999–6009, 2017.

[2] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[3] Anthony Brohan et al. RT-1: robotics transformer for real-world control at scale. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023.

[4] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

[5] Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[6] Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3756–3762, 2017.

[7] Jernej Kos and Dawn Song. Delving into adversarial attacks on deep policies. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

[8] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

[9] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020.

[10] Tuomas Oikarinen, Wang Zhang, Alexandre Megretski, Luca Daniel, and Tsui-Wei Weng. Robust deep reinforcement learning through adversarial loss. *Advances in Neural Information Processing Systems*, 34:26156–26167, 2021.

[11] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33:19884–19895, 2020.

[12] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pages 5639–5650. PMLR, 2020.

[13] Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International conference on learning representations*, 2020.

[14] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114. PMLR, 2020.