オフライン強化学習におけるデータ拡張によるAtariゲームAIの性能向上

髙野 剛志[1,a)]　計良 宥志[2,b)]　川本 一彦[2,c)]

概要：本研究では，オフライン強化学習の一手法である Decision Transformer を用い，Atari ゲームの画面にガウスノイズ等を付加するデータ拡張を通じて，エージェントの性能が向上することを示す．このデータ拡張は通常，ノイズに対するロバスト性の向上を目的としているが，本研究ではクリーンなゲーム環境においても性能が向上することを発見した．特に，ノイズデータとクリーンデータを組み合わせて訓練することで，クリーンデータあるいはノイズデータのみを用いた場合よりも高いスコアが得られることを実験的に示す．これらの結果は，ノイズ付与によるデータの多様性が Decision Transformer の性能向上に有益であることを示唆している．

Improving Atari Game AI Performance via Data Augmentation in Offline Reinforcement Learning

Tsuyoshi TAKANO[1,a)]　Hiroshi KERA[2,b)]　Kazuhiko KAWAMOTO[2,c)]

Abstract: In this study, we demonstrate that using the Decision Transformer, an offline reinforcement learning method, together with data augmentation by adding noise to Atari game screens, improves the performance of the agent. This data augmentation is typically aimed at improving robustness against noise, but in this study, we find that performance also improves in clean game environments. In particular, training with a combination of noisy and clean data results in higher scores compared to using only clean data or only noisy data. These results suggest that the diversity introduced by noise is beneficial for improving the performance of the Decision Transformer.

1. はじめに

オフライン強化学習は，エージェントが環境と相互作用することなく，事前に収集したデータセットを用いて方策を学習する手法である．この手法は，特に人間プレイヤーのログデータを活用することで，ゲーム AI への応用が期待されている．近年，このオフライン強化学習に対する新しいアプローチとして，Decision Transformer が提案されている [1]．Decision Transformer は，Transformer のアテンション機構 [2] を活用し，系列データから次の行動を予測するタスクを解決することで，オフライン強化学習を実現する．

一方，強化学習におけるデータ拡張手法は，エージェントの汎化性能を向上させるために広く研究されている．RAD [3] や CURL [4] といった手法では，画像の回転やクロッピング，対照学習などを用いたデータ拡張が行われており，クリーン環境下での性能向上を目的としている．これらのデータ拡張手法は，画像分類タスクにおいては効果的であることが多い．しかし，Atari ゲームにおいては必ずしも有効とは言えない．例えば，Pong タスクにおいて画像を回転させるデータ拡張を行うと，ゲームのルールに反してエージェントが行動を学習する可能性がある．このよ

[1] 千葉大学大学院 融合理工学府 数学情報科学専攻
Graduate School of Science and Engineering, Chiba University
[2] 千葉大学大学院 情報学研究院
Graduate School of Informatics, Chiba University
[a)] tsuyoshi.takano@chiba-u.jp
[b)] kera@chiba-u.jp
[c)] kawa@faculty.chiba-u.jp

うに，Atari ゲームでは環境の性質により，従来のデータ拡張手法が性能低下を引き起こす恐れがある．Yang ら [5] や高野ら [6] は，ノイズデータに対するロバスト性に焦点を当てた手法を提案している．これらの研究では，ノイズに対してモデルの耐性を向上させることが主な目的であり，クリーンな環境における性能については明らかではない．

そこで，本研究では，Atari ゲーム環境において，ゲーム画面にノイズを付加するデータ拡張により，Decision Transformer の性能向上を目指す．ノイズは画像の内容や構造を大きく変えないため，環境の整合性を保ちながらデータの多様性を提供することが可能である．具体的には，ガウスノイズなどの 4 種類のノイズを訓練データに付加し，それがモデルの学習に与える影響を評価する．一般に，深層学習による画像分類では，データ拡張によりノイズに対するロバスト性が向上する反面，正常時の性能が劣化するというトレードオフが存在する．しかし，本研究では，Decision Transformer においてクリーン環境下での性能も向上することを示す．

2. 手法

本研究では，Decision Transformer を用いて，Atari のゲームタスクに対する性能向上を検証する．ノイズデータとクリーンデータを異なる比率で組み合わせたデータセットを使用し，データの多様性を変化させながら，エージェントの性能を評価する．データ拡張には，ガウスノイズ，ショットノイズ，インパルスノイズ，スペックルノイズ [7] の 4 種類を観測データに付加する．

2.1 Decision Transformer

Decision Transformer は，観測データ s_t，行動データ a_t，および将来報酬 $\hat{R}_t$ を入力として次の行動を予測する系列モデルである．トークン系列は以下のように定義される．

$$\tau = \left(\hat{R}_1, s_1, a_1, \hat{R}_2, s_2, a_2, \ldots, \hat{R}_T, s_T, a_T\right). \quad (1)$$

ここで，$\hat{R}_t$ は時刻 t から終端までの累積報酬 $\hat{R}_t = \sum_{t'=t}^{T} r_{t'}$ である．Decision Transformer は，このトークン系列を Transformer に入力し，因果マスクを用いて未来の情報を利用せずに次の行動 a_{t+1} を予測する．

2.2 ノイズ付加によるデータ拡張

Decision Transformer の訓練には，Atari のゲームタスクから収集した DQN-Replay データセット [8] を使用する．このデータセットには，エージェントの観測データ，行動，報酬の履歴が含まれている．このデータセットに対して，以下の 4 種類のノイズを観測データに付加する．各ノイズデータに関して，0 %から 100 %までの比率でクリーンデータと組み合わせ，複数の訓練データセットを構築する．例えば，ノイズ比率が 40 %の場合，データセットの 40 %がノイズを付加したデータ，60 %がクリーンデータから構成される．

ガウスノイズ: 各画素 x にガウスノイズを追加する．

$$x_{\text{noisy}} = \text{clip}(x + v, 0, 1) \quad (2)$$

ここで，$N(0, \sigma)$ は，平均 0，標準偏差 $\sigma = 0.08$ のガウス分布で，v はこのガウス分布からのサンプルである．また，clip(x, 0, 1) は画素値を 0 から 1 の範囲にクリップする関数とする．

ショットノイズ: ショットノイズはポアソン分布に従い，画素値 x に対してノイズを加える．

$$x_{\text{noisy}} = \text{clip}\left(\frac{x'}{60}, 0, 1\right) \quad (3)$$

ここで，x' は，平均 $\lambda = 60x$ のポアソン分布 $P(\lambda)$ からのサンプルである．

インパルスノイズ: ランダムに選択した画素値を 0 または 1 に置換する．

$$x_{\text{noisy}} = \begin{cases} 0 & \text{with probability } 0.03 \\ 1 & \text{with probability } 0.03 \\ x & \text{otherwise} \end{cases} \quad (4)$$

ここで，0.03 はインパルスノイズが適用される確率を表す．

スペックルノイズ: 各画素値に乗算的なノイズを付加する．ここで，v は $v \sim N(0, 0.15)$ に従ってサンプリングされるノイズ項を示し，元の画像 x に対して乗算された後に加えられる．

$$x_{\text{noisy}} = \text{clip}(x + xv, 0, 1) \quad (5)$$

2.3 性能評価

データ拡張したデータセットを使用し，各ノイズ比率 (0 %, 20 %, 40 %, 60 %, 80 %, 100 %) について，Decision Transformer を訓練する．訓練に用いたハイパーパラメータを表 1 に示す．訓練データはランダムサンプリングする．

性能評価では，クリーンな観測データを用いたゲーム環境で，訓練後のエージェントの平均ゲームスコアを評価する．

3. 実験

4 種類の Atari ゲームタスクを対象とし，クリーンな評価環境において Decision Transformer の性能を評価する．評価の目的は，ガウシアン，インパルス，ショット，スペックルの 4 種類のノイズを訓練データに適度に混ぜることで，クリーンデータのみを使用する場合と比較して，エージェントの性能がどのように変化するかを明らかにすることである．

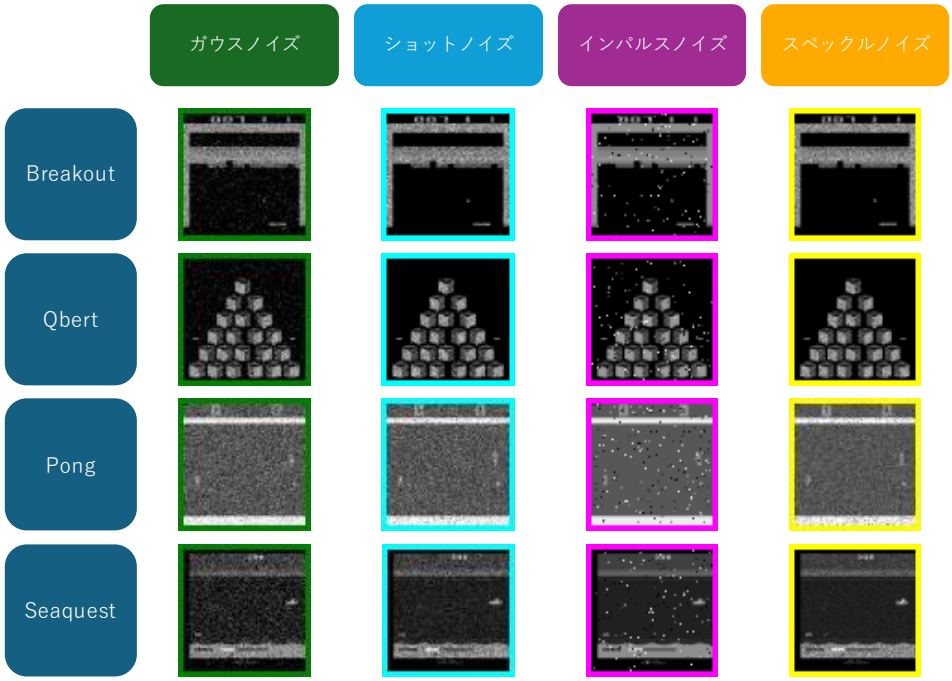

図 1　データ拡張の例

表 1　ハイパーパラメータ

Hyperparameter	Value
Batch size	128
Context length K	50 (Pong) 30 (Breakout) 30 (Qbert) 30 (Seaquest)
Return-to-go conditioning	90 (Breakout) 14000 (Qbert) 20 (Pong) 1150 (Seaquest)
Max epochs	5 (all)
Replay buffer capacity	500000 (all)

3.1　実験結果

4 種類のゲームに対して，クリーンデータとノイズデータの比率に対する平均スコアを表 2 から表 5 に示す．表 2 から表 5 は，1 エポック毎に 5 回評価を行い，5 エポック分の 25 回評価を平均したスコアを使用している．以下では，各ゲームに関する結果をまとめる．

Breakout 表 2 から，ショットノイズを除く 3 種類のノイズで，ノイズ比率 0%（クリーンな環境での訓練）と比べて，平均スコアが向上するノイズ比率が存在している．ガウスノイズやインパルスノイズではノイズ比率 20%，スペックルノイズではノイズ比率 80%のときに，最もスコアが高くなっている．とくに，ガウスノイズのノイズ比率 20%のときに，平均スコアが 67.32 と全体で最も高いスコアを達成している．

Qbert 表 3 から，すべてのノイズについて，ノイズ比率 0%と比べて，平均スコアが向上するノイズ比率が存在している．他のゲームと比べて，高いノイズ比率（80%あるいは 100%）でのデータ拡張のときに，最も高いスコアになっている．とくに，インパルスノイズのノイズ比率 100%のときに、平均スコアが 7373.0 と全体で最も高くなっている．全体的にスコアの向上も大きく，4 種類のゲームのなかで，ノイズデータ拡張が最も有効であったといえる．

Pong 表 4 から，すべてのノイズについて，ノイズ比率 0%と比べて，平均スコアが向上するノイズ比率が存在している．各ノイズでゲームスコアが高くなるノイズ比率は 20%から 60%であり，Breakout と Qbert の間に最適なノイズ比率があることを示唆している．全体では，ガウスノイズ 40%のときに，平均スコアが 15.52 と最も高くなっている．

Seaquest 表 5 から，ショットノイズとスペックルノイズでスコア向上が見られるものの，その向上は小さい．さらに，ガウスノイズとインパルノイズでは，ノイズ比率 0%のときに最もスコアが高くなっている．したがって，4 種類のゲームのなかでは，ノイズデータ拡張の効果が最も小さい結果になった．

3.2　考察

ゲームによって最適なノイズ比率が異なることが確認された．Qbert では，ノイズ比率が高い（80〜100%）ときにスコアが向上し，データの多様性がエージェントの学習に大きく貢献していることが示唆された．一方で，Breakout では，低いノイズ比率（20%）で最も高いスコアが得られ，

表 2　ノイズ種別ごとの Breakout の平均スコア

ノイズ比率	ガウスノイズ	ショットノイズ	インパルスノイズ	スペックルノイズ
0%	55.36 ± 49.47	**69.80 ± 72.13**	37.72 ± 19.72	56.44 ± 53.83
20%	**67.32 ± 74.26**	32.80 ± 25.86	**38.76 ± 19.40**	45.52 ± 30.37
40%	34.68 ± 20.99	53.76 ± 37.10	27.48 ± 44.48	48.64 ± 52.92
60%	33.64 ± 21.55	47.68 ± 26.48	8.92 ± 5.61	51.92 ± 29.68
80%	54.56 ± 53.34	44.32 ± 23.75	5.00 ± 5.01	**57.88 ± 48.06**
100%	5.60 ± 3.65	49.44 ± 45.55	3.80 ± 2.84	54.44 ± 42.50

表 3　ノイズ種別ごとの Qbert の平均スコア

ノイズ比率	ガウスノイズ	ショットノイズ	インパルスノイズ	スペックルノイズ
0%	2813.0 ± 3166.7	2219.0 ± 3191.5	4121.0 ± 2823.9	3308.0 ± 3203.2
20%	2510.0 ± 2606.6	2802.0 ± 3273.7	6283.0 ± 3792.3	1384.0 ± 1609.0
40%	2144.0 ± 2761.7	1079.0 ± 1465.4	2618.0 ± 3309.1	3635.0 ± 3568.8
60%	2747.0 ± 2848.8	1981.0 ± 2298.0	4652.0 ± 3349.7	2376.0 ± 3220.2
80%	4695.0 ± 4294.9	**4466.0 ± 4697.6**	1860.0 ± 2108.3	4089.0 ± 3830.8
100%	**6254.0 ± 3768.3**	530.0 ± 715.6	**7373.0 ± 4604.4**	**4841.0 ± 3824.5**

表 4　ノイズ種別ごとの Pong の平均スコア

ノイズ比率	ガウスノイズ	ショットノイズ	インパルスノイズ	スペックルノイズ
0%	10.16 ± 11.25	7.64 ± 15.43	7.92 ± 15.00	8.84 ± 13.68
20%	14.84 ± 6.63	13.04 ± 8.71	**12.12 ± 10.08**	1.72 ± 16.29
40%	**15.52 ± 6.08**	7.60 ± 14.80	10.16 ± 13.99	11.64 ± 10.87
60%	11.40 ± 10.05	**15.08 ± 7.39**	9.52 ± 13.55	**14.84 ± 5.65**
80%	11.80 ± 11.80	9.84 ± 13.74	8.92 ± 14.96	8.76 ± 14.91
100%	7.16 ± 13.52	11.00 ± 9.67	6.16 ± 12.46	12.32 ± 9.73

表 5　ノイズ種別ごとの Seaquest の平均スコア

ノイズ比率	ガウスノイズ	ショットノイズ	インパルスノイズ	スペックルノイズ
0%	**1030.4 ± 340.1**	995.2 ± 356.0	**917.6 ± 340.3**	923.2 ± 421.8
20%	910.4 ± 259.1	809.6 ± 372.9	904.0 ± 447.7	936.0 ± 355.3
40%	1004.8 ± 454.9	768.0 ± 392.8	796.0 ± 390.7	**1011.2 ± 404.6**
60%	954.4 ± 421.4	**1080.8 ± 285.6**	746.4 ± 384.8	968.8 ± 345.1
80%	792.0 ± 275.6	762.4 ± 367.6	759.2 ± 366.0	782.4 ± 416.2
100%	787.2 ± 369.1	994.4 ± 332.3	677.6 ± 321.9	891.2 ± 395.6

高いノイズ比率では逆に性能が低下していることが分かった．Pong では，適度なノイズ比率（20〜60%）が最も効果的であり，過剰なノイズ付与は逆効果となることが示された．また，Seaquest では，ノイズの付加が性能向上の妨げになる例も確認され，すべてのノイズで有効とは限らないことがわかった．

総じて，ゲームに適したノイズの種類や比率の選択が重要であり，適切に調整することで効果的な学習を促進する可能性がある．今後の課題としては，ゲームに最適なノイズの種類や比率を自動で調整する手法の開発や，異なる環境への応用可能性を検討する必要がある．

4. おわりに

本研究では，Atari ゲームにおいて，適度にノイズを付加したデータが Decision Transformer の性能向上に寄与することを示した．ノイズによるデータ拡張が訓練データセットの多様性を高め，より広範な観測に対するエージェントの適応能力を向上させたと考えられる．一方で，ゲームによって最適なノイズ種類や比率は異なっていることも確認され，ノイズ比率の調整が学習の鍵であることが示唆された．

今後の研究としては，異なるノイズ手法やデータ拡張手法のさらなる検証が求められる．例えば，特に，ガウスノイズやショットノイズ以外の手法や，回転・反転といった他の一般的なデータ拡張技術の効果についても調査することで，より幅広い環境での有効性が確認されることが期待される．さらに，ロボット制御など，異なるタスクへの適用も重要な課題であり，ノイズを用いたデータ拡張手法の応用範囲を広げるための研究が求められる．

謝辞

本研究は JSPS 科研費 JP23K24914 の助成を受けたものである．

参考文献

[1] Chen, Lili, Lu, Kevin, Rajeswaran, Aravind, Lee, Kimin, Grover, Aditya, Laskin, Misha, Abbeel, Pieter, Srinivas, Aravind, and Mordatch, Igor. Decision transformer: Reinforcement learning via sequence modeling. Advances in Neural Information Processing Systems, 34:15084–15097, 2021.

[2] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, and Polosukhin, Illia. Attention is all you need. Advances in Neural Information Processing Systems, 30:5999–6009, 2017.

[3] Laskin, Michael, Srinivas, Aravind, and Abbeel, Pieter. Reinforcement learning with augmented data. Advances in Neural Information Processing Systems, 33:19884–19895, 2020.

[4] Srinivas, Aravind, Laskin, Michael, and Abbeel, Pieter. CURL: Contrastive unsupervised representations for reinforcement learning. Proceedings of the 37th International Conference on Machine Learning (ICML), pages 5549–5559, 2020.

[5] Yang, Rui, Zhong, Han, Xu, Jiawei, Zhang, Amy, Zhang, Chongjie, Han, Lei, and Zhang, Tong. Towards Robust Offline Reinforcement Learning under Diverse Data Corruption. International Conference on Learning Representations , 2024.

[6] 高野 剛志, 計良 有志, 川本 一彦. Atari ゲームに対する Transformer ベース強化学習のロバスト性検証. 2024 年度人工知能学会全国大会（第 38 回）, 2024.

[7] Hendrycks, Dan, and Dietterich, Thomas. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. International Conference on Learning Representations, 2018.

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