Marltas: ヘッドレスブラウザを用いた 強化学習のためのブラウザゲーム実行環境

佐々木 一磨 †, 大垣 慶介, 小田桐 優理
†kazuma_sasaki@dwango.co.jp
株式会社ドワンゴ

2019/7/30

概要

深層強化学習のベンチマークとして様々なビデオゲームのタスクセットが提案されている．しかしタスクセットの実装にはゲームエンジンへの理解など技術的なコストが高い．そこで我々はブラウザゲームに着目し，タスクを規定するゲームロジックとゲーム実行環境の実装を分離することで容易にタスク追加可能な開発環境”Marltas”を提案する．実装例として複数のゲームについて人間と深層強化学習の評価実験を行い，ベンチマークを示した．

1 はじめに

近年，深層強化学習の研究が注目されている．強化学習は環境から得られる状態から事前に設定した状態の良さを反映した報酬を最大化する方策関数を最適化することを目的とする．この方策関数を深層学習モデルで近似することで従来では現実的ではなかった画像などの多次元情報を状態として直接扱うことができるようになった [1]．現在ではその高い学習性能と入力の汎用性からロボティクスを中心にして様々なタスクへの応用が期待されている．

強化学習では方策関数をもつエージェントが環境とインタラクトすることで学習データを収集する必要があり，その環境としてビデオゲームがしばしば用いられる．特にビデオゲームはプレイの達成度を評価指標として利用できるため，ベンチマークのタスクとして利用されている [2]．また対人対戦 [3], 3D シューティング [4, 5, 6]，マルチエージェント [7, 5] などの発展的なタスクとしてもゲームが採用されている．

ビデオゲームによる強化学習実験はゲーム内容を変更することで様々タスクが実現可能であるのに対して，実装を行うための技術的なコストが高い．タスクセット開発にはゲーム画面のレンダリング，ゲームエンジン，人間と強化学習エージェントのためのなどの複数の機能が要求される．ATARI のタスクセット [2] は複数の環境を実装しているが，ゲームの表現能力が限られている．ML-Agent[8] はタスク実装のためのゲーム開発環境 Unity のプラグインで，強化学習のための機能を提供しているが，環境の振る舞いを決めるゲームロジックの実装を必要とする．

そこで我々はタスクの実装をゲームエンジンとゲームロジックに分離することで用意にタスク追加が可能な強化学習学習環境 Multi-process Asynchronous Reinforcement Learning based Game Tactics Acquisition(Marltas) を提案する．Marltas はヘッドレスブラウザ上で動作するゲームと強化学習エージェントとのとして機能する．エージェントの取る行動はブラウザの操作，状態はゲーム画面画像と

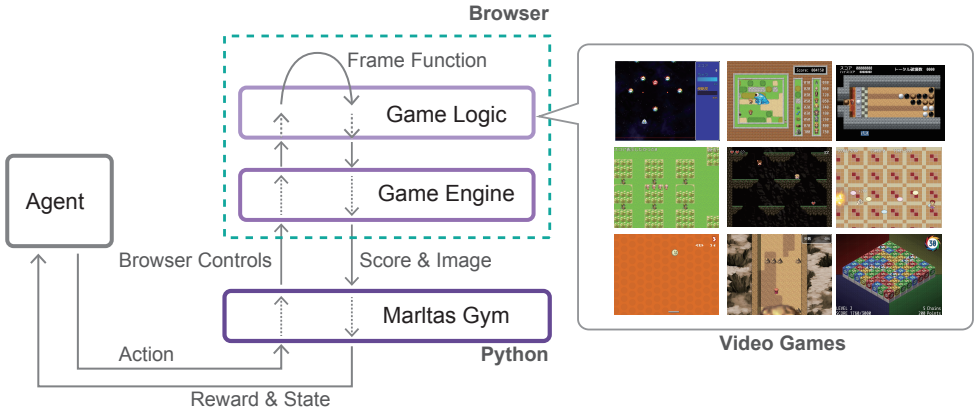

図 1: Marltas の構成

して共通化されており，そのためタスクの追加を行うためにゲームエンジンやの実装を変更する必要がない．

本稿の構成は以下のようになっている．2 章では Marltas の設計指針であるゲームエンジンとロジックの分離と，それによって得られる利点について述べる．3 章では実際にゲームエンジンとして「RPG ツクール MV」[9] を利用した Marltas の深層強化学習実験を行う．最後に 4 章では本稿のまとめと今後の展望について述べる．

2 Marltas の設計指針

図 1 に Marltas の構成を示す．強化学習環境は Marltas Gym, Game Engine, Game Logic と呼ばれる 3 つの層を通してブラウザゲームを実行する．Martas Gym 層はエージェントであり，OpenAI Gym[10] として振る舞う．Marltas Gym 層はエージェントから出力された行動ベクトルによってブラウザを操作し，ゲームを進行させる．そして更新されたゲーム画面とスコアに基づいた新しい状態と報酬をエージェントへと返す．ゲームの機能は画面レンダリングなどのゲームロジック共通の機能を持つ Game Engine 層，タスク固有の状態遷移の機能に対応する Game Logic 層の 2 つによって実装されている．また，学習実験が CUI のみが利用可能なサーバー上で実行されることを想定し，ブラウザとして GUI 機能を持たないヘッドレスブラウザを利用する．

Marltas ではゲームロジック共有の機能が提供されている．そのため，別のタスクに対する実装を行う際には通信やレンダリングなどの低レイヤ処理の変更は不要であり，タスク依存である Game Logic 層に関する部分のみを変更すればよい：

- ゲームプレイ開始までの操作（例：ゲーム難易度を選択してスタートボタンを押す）
- エージェントから与えられる行動ベクトルとゲーム操作の対応
- スコアに基づく報酬関数
- ゲームオーバー後のリセット操作

3 実験

Marltas が強化学習研究で利用可能であることを確かめるため，ゲームエンジンとして「RPG ツクール MV」を利用した評価実験を行った．

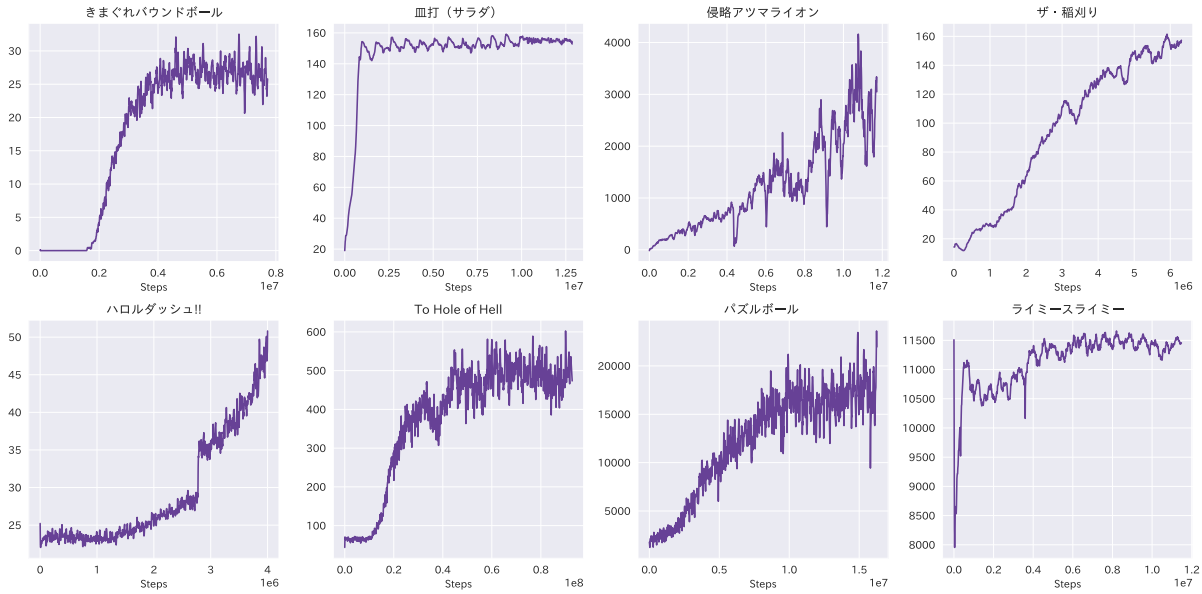

図 2: DQN の学習曲線．横軸: 学習に利用されたエピソードの累積ステップ数．縦軸: エピソード報酬の平均値

3.1 実装

「RPG ツクール MV」はロールプレイングゲームを中心としたゲームの開発環境であり，ブラウザ上で実行可能なコアスクリプトが Web 上で公開されている [9].「RPG アツマール」[11] ではこのエンジンを使って開発されたゲームを配信しており，本実験ではこの中から 8 種類のゲームを利用した．以下に各ゲームの内容を簡易に説明する：

- きまぐれバウンドボール [12]: 自由落下するボールを落とさないように打ち返し続けるゲーム．マウスによる操作を行う．
- 皿打 (サラダ) [13]: 画面内のキャラクターが射出するアイテム（皿）を打ち落とし続けるゲーム．マウスによる操作を行う．
- ザ・稲刈り [14]: 制限時間以内にマップ内を移動してアイテムを集めるゲーム．キーボードによる上下左右の操作．
- ハロルダッシュ!! [15]: 強制スクロールで進むキャラクターの前方から迫ってくる岩を避け続けるゲーム．キーボードによる左右の操作．
- To Hole of Hell [16]: 強制スクロールで画面内の島となっているステージをつたって降りていくゲーム．キーボードによる左右の操作．
- ライミースライミー [17]: 制限時間以内にマップ内を移動してアイテムを集めるゲーム．マウスによる操作．
- パズルボール [18]: 画面内の石に対して同色の石をぶつけて消していくゲーム．キーボードによる上下左右の操作．
- 侵略アツマライオン [19]: 前方からスクロールしてくる敵を倒していくシューティングゲーム．キーボードによる操作．

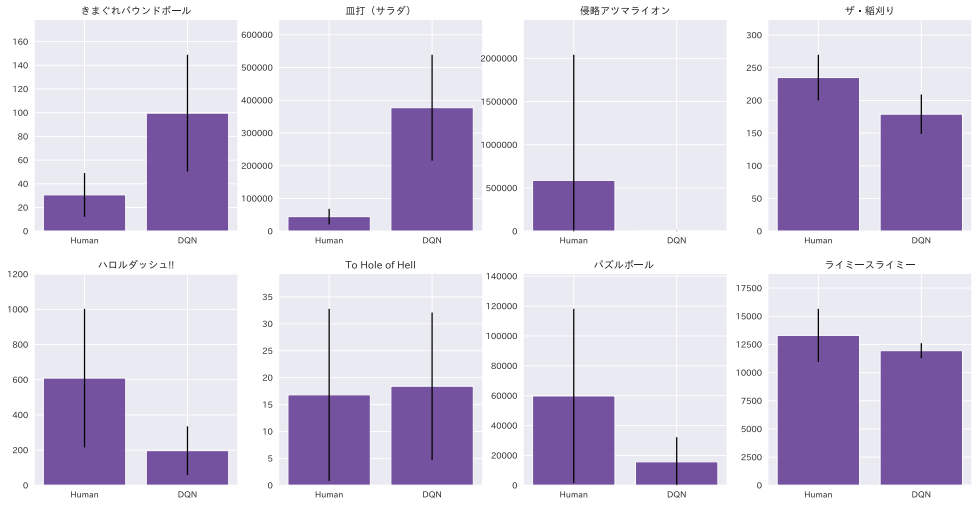

図 3: 各ゲームにおけるスコアの比較．縦軸：エピソードの報酬平均と標準偏差.「侵略アツマライオン」の DQN スコアは人間スコアと比べ非常に小さい値（平均 2107.8）になったため省略した．

表 1: Q-Network の構成．N は行動ベクトルの大きさを表す.

#	Layer	Param.
1	Convolution2D	filter:8 × 8, stride:4
2	Convolution2D	filter:8 × 8, stride:2
3	Convolution2D	filter:4 × 4, stride:2
4	Convolution2D	filter:8 × 8, stride:1
5	Convolution2D	N

エージェントへの入力である状態はゲーム画面の 128 × 128 ピクセルの画像を用いて，4 ステップ分をチャンネル方向に結合した 128 × 128 × 12 の行列とした．行動はキーボードとマウスによるブラウザ操作を対象タスク（ゲーム）に応じて離散化したベクトルを用いた．報酬はゲームから得られるスコアの差分を基本的に用いた．スコアが明示的に与えられないゲームの場合には達成度を示す別のゲーム状態を利用した．また一部のゲームについては学習の効率化を行うために補助報酬の追加と早期打ち切りを行った．

強化学習手法として Deep Q-Network（DQN, [1]）を採用した．学習方法として Horgan らによって提案された APE-X を用いた [20]．学習計算の実行環境はマルチノードにおける分散処理実行環境である RLlib[21] を用いた．学習パラメーターは RLlib のデフォルト値を元にして調整を行った．特に，提案された APE-X の損失関数では複数ステップの TD 誤差を計算していたが，今回は 1 ステップのみの誤差を利用した．表 1 に Q-Network モデルの構造を示す．モデルは 4 層の畳み込み層による Convolutional Neural Network[22] とし，活性関数は ReLu[23] を用いた．

学習の達成度を評価するために学習後の DQN モデルを利用して各ゲームについて 30 回試行を行いスコアを記録した．さらに DQN との比較として人間のスコアを評価した．人間のスコアは個々のゲームについて総プレイ時間が 10 時間以内の初心者，20 代〜30 台の男女 5 名がそれぞれ 5 回プレイしたときのスコアとした．ただし,「パズルボール」については初心者でない被験者 1 名分のスコ

アを除外した．

3.2 実験結果と考察

図 2 に個々のゲームの学習時のエピソードの報酬平均の推移を示す．横軸は学習時のエピソードのステップ数である．また，図 3 にスコア結果の比較を示す.「きまぐれバウンドボール」「皿打（サラダ）」では人間に対して DQN が高いスコアを示した.「To Hole of Hell」「ライミースライミー」「ザ・稲刈り」ではほぼ人間と同様のスコアとなった．その他の「侵略アツマライオン」「ハロルダッシュ!!」「パズルボール」は人間のスコアを下回った．

8 種類のゲームについての結果，内 5 種類については人間を上回るかほぼ同等のスコアを示した．これは Marltas が強化学習フレームワークとして有用であることを示している．人間のスコアを下回った 3 種類のゲームについては強化学習の手法を評価するための新たな課題となりうると考えられる．

4 結論

本稿ではブラウザを用いた強化学習のためのビデオゲーム実行環境 Marltas を提案した．強化学習研究ではビデオゲームによるタスクがベンチマークとして用いられているが，個々のタスクに依存したゲームエンジン，実行環境を用意する必要がありタスクの追加が容易ではなかった．そこで我々はブラウザゲームに着目し，ゲームエンジンとゲームロジックを分離しタスク共有の機能をまとめる実行環境の実装方法を提案した．具体例としてゲームエンジン RPG ツクール MV を利用した環境を実装し，実際に強化学習実験を行うことで Marltas が実験環境として利用可能であることを示した．

今後の展望として対応ゲームの拡大，任意のゲーム状態へのリセットの対応があげられる．RPG アツマールでは RPG ゲームが多く提供されている．RPG ゲームは強化学習問題として捉えた場合，マップ内での移動，メニュー画面の操作，戦略的な行動選択などの複合的かつ長期に渡るエピソードであり，ϵ-greedy のようなランダムの行動選択に頼る探索では限界があるだろう．Ecoffet らによって提案されたゲームの途中状態からの開始を利用した探索手法 [24] には任意状態へのリセットは不可欠である．

参考文献

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, pp. 529–533, feb 2015.

[2] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, “The arcade learning environment: An evaluation platform for general agents,” *Journal of Artificial Intelligence Research*, vol. 47, pp. 253–279, jun 2013.

[3] O. Vinyals, I. Babuschkin, J. Chung, M. Mathieu, M. Jaderberg, W. M. Czarnecki, A. Dudzik, A. Huang, P. Georgiev, R. Powell, T. Ewalds, D. Horgan, M. Kroiss, I. Danihelka, J. Agapiou, J. Oh, V. Dalibard, D. Choi, L. Sifre, Y. Sulsky, S. Vezhnevets, J. Molloy, T. Cai, D. Budden, T. Paine, C. Gulcehre, Z. Wang, T. Pfaff, T. Pohlen, Y. Wu, D. Yogatama,

J. Cohen, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, C. Apps, K. Kavukcuoglu, D. Hassabis, and D. Silver, "AlphaStar: Mastering the Real-Time Strategy Game StarCraft II." https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/, 2019.

[4] M. Kempka, M. Wydmuch, G. Runc, J. Toczek, and W. Jakowski, "Vizdoom: A doom-based ai research platform for visual reinforcement learning," in *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pp. 1–8, Sep. 2016.

[5] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman, N. Sonnerat, T. Green, L. Deason, J. Z. Leibo, D. Silver, D. Hassabis, K. Kavukcuoglu, and T. Graepel, "Human-level performance in first-person multiplayer games with population-based deep reinforcement learning," 2018.

[6] C. Beattie, J. Z. Leibo, D. Teplyashin, T. Ward, M. Wainwright, H. Kttler, A. Lefrancq, S. Green, V. Valds, A. Sadik, J. Schrittwieser, K. Anderson, S. York, M. Cant, A. Cain, A. Bolton, S. Gaffney, H. King, D. Hassabis, S. Legg, and S. Petersen, "Deepmind lab," 2016.

[7] J. Suarez, Y. Du, P. Isola, and I. Mordatch, "Neural mmo: A massively multiagent game environment for training and evaluating intelligent agents," 2019.

[8] A. Juliani, V.-P. Berges, E. Vckay, Y. Gao, H. Henry, M. Mattar, and D. Lange, "Unity: A general platform for intelligent agents," 2018.

[9] KADOKAWA CORPORATION and Y. OJIMA, "Rpg maker mv." http://www.rpgmakerweb.com/products/programs/rpg-maker-mv, 2017.

[10] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016.

[11] DWANGO Co., Ltd., "Rpg アツマール." https://game.nicovideo.jp/atsumaru/.

[12] りばりー, "きまぐれバウンドボール." https://game.nicovideo.jp/atsumaru/games/gm6285.

[13] うどんぱ, "皿打 (サラダ)." https://game.nicovideo.jp/atsumaru/games/gm7506.

[14] NEZU, "ザ・稲刈り." https://game.nicovideo.jp/atsumaru /games/gm7972.

[15] 微熱, "ハロルダッシュ!!." https://game.nicovideo.jp/atsumaru/games/gm2525.

[16] あおいたく, "To hole of hell." https://game.nicovideo.jp/atsumaru/games/gm6577.

[17] きつねうどん, "ライミースライミー." https://game.nicovideo.jp/atsumaru/games/gm2147.

[18] 北. 舞楽, "パズルボール." https://game.nicovideo.jp/atsumaru/games/gm3788.

[19] THE ファンキー, "侵略アツマライオン." https://game.nicovideo.jp/atsumaru/games/gm3144.

[20] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. van Hasselt, and D. Silver, "Distributed Prioritized Experience Replay," in *nternational Conference on Learning Representations*, pp. 1–19, 2018.

[21] E. Liang, R. Liaw, P. Moritz, R. Nishihara, R. Fox, K. Goldberg, J. E. Gonzalez, M. I. Jordan, and I. Stoica, "RLlib: Abstractions for Distributed Reinforcement Learning," *International Conference on Machine Learning*, dec 2017.

[22] A. Krizhevsky, I. Sutskever, and H. Geoffrey E., "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 1–9, 2012.

[23] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (G. Gordon, D. Dunson, and M. Dudk, eds.), vol. 15 of *Proceedings of Machine Learning Research*, (Fort Lauderdale, FL, USA), pp. 315–323, PMLR, 11–13 Apr 2011.

[24] A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune, "Montezuma ' s revenge solved by go-explore, a new algorithm for hard-exploration problems (sets records on pitfall, too)." https://eng.uber.com/go-explore/, 2018.