

MarioAIにおけるデフォルト行動を前提とした新たな効率的な学習方法の提案

平野 笙太郎^{1,a)} 渡邊 真也^{†1,b)} 半田 久志^{†2,c)}

概要: 本研究では、代表的な学習用ベンチマークである MarioAI に対する新たな効率的な学習アプローチの提案とその有効性の検証を行う。提案するアプローチでは、問題に対する事前知識として概ねこういった行動が良いという標準的行動が存在する場合を想定し、それが上手く機能しない場合についての学習を行う。そのため、多くの場合においてそれが当てはまる問題に対しては、学習すべき量が大幅に減少するため、高い学習効率を期待することができる。本研究では、MarioAI に対して提案するアプローチを実装し、ニューラルネットワークや進化計算を利用した場合との比較を行い、学習効率および学習の質についての検証を行った。

Proposal of New Efficient Learning Method Considering Standard Action in MarioAI

Abstract: In this paper, a new efficient learning approach considering standard action for MarioAI is proposed. In the proposed learning approach, standard (default) action means more or less best action in most cases and MarioAI could be premised on this default action. This approach tries to find situations that standard action does not work well and learn which action is suitable for this situation. Since these situations are much more limited than the number of all situations in MarioAI, the proposed approach can be expected to be efficient and effective. In this paper, the effectiveness of the proposed approach was verified by comparing the performance of the neural network and evolutionary learning in MarioAI benchmark problem.

1. はじめに

デジタルゲームプログラミングの国際学会である IEEE Symposium on Computational Intelligence and Games では、毎年様々なデジタルゲームを題材にしたコンペティションが 2009 年から開催されており [1]、その中の興味深い取り組みの 1 つに Mario AI Benchmark を用いた Mario AI Competition がある [2]。

ゲーム AI には、ゲームを用いて AI を改善する研究と

AI を用いてゲームを改善する研究の 2 種類が存在するが、Mario AI は前者にカテゴライズされる。これまで様々なアルゴリズムが Mario AI を含むリアルタイムゲームに適用されており、優れた AI アルゴリズムが多数開発されている [3]。また、単に優れたゲームスコアをたたきだす AI ではなく、人間のプレイに似た行動を学習する AI についても研究が進められており、人間とより協調することのできる AI として期待されている。

このように、Mario AI を含むリアルタイムゲームを用いたゲーム AI アルゴリズムの研究は非常に盛んに行われており、DBM(Deep Boltzmann Machine) 等のニューラルネットワークを用いたアプローチ [4] や、進化的アルゴリズムを用いたアプローチ [5] など、様々な手法による興味深い成果が報告されている。

本研究では、MarioAI における新たな学習アプローチとして、標準的行動（以下、定石）を想定し、定石の行動ではうまく行かない場合のみ学習を行う方法を提案する。MarioAI を含むいくつかのゲームでは、多くの場合におい

¹ 室蘭工業大学大学院
The Graduate School of Engineering, Muroran Institute of Technology

^{†1} 室蘭工業大学 しくみ解明系領域
Presently with College of Information and Systems, Muroran Institute of Technology

^{†2} 近畿大学 理工学部 情報学科
Presently with School of Science and Technology, Kinki University

a) 19043058@mmm.muroran-it.ac.jp

b) sin@csse.muroran-it.ac.jp

c) handa@info.kindai.ac.jp

て概ねこういった行動が良いという定石が存在する場合が少なくなく、提案手法はその定石を最大限利用することで、学習すべき場面を限定し、学習の効率化を図ろうとしている。

提案手法は、定石が通用しない（タブー）の時のみ学習を行うため、ある種のタブー学習と捉えることができる。そこで、本論文では便宜上、提案手法を TLVET(Taboo Learning Visible from Established Tactic) と呼ぶ。

Mario AI における最大のタブーは“死”と“停止”であるため、TLVET はそういったタブーの状態を回避する為に、そのタブー状態に至る直前の行動を学習することで、タブー状態の回避を試みる。そのため、TLVET は定石が強く通用する場合の問題に対して特に有効的な結果を示すと考えており、逆に定石がほとんど通用しない、もしくは定義できない問題に対してはうまく機能しないと思われる。

本研究では、MarioAI における TLVET の有用性を検証するため、ニューラルネットワークの1種である SRN(Simple Recurrent Network), MLP(Multilayer perceptron) および進化計算を学習に用いた手法などとの比較実験を行い、それぞれの手法による結果の分析を行った。

以下、本稿における構成について述べる。まず2章では、取り扱うベンチマークである Mario AI について、3章では Mario AI に関する研究について概説する。その上で、4章では提案するデフォルト行動を前提とした手法の詳細について説明する。5章において提案手法の検証実験を行い、6章にまとめを述べる。

2. Mario AI

本研究では、Mario AI Championship サイトにより提供されるシミュレーションプログラムを用いた [2]。プログラムは、いわゆる任天堂のスーパーマリオブラザーズをシミュレートした競技であり、表1に示す4つの部門に分かれている。Mario AI Benchmark ライブラリでは、この Level Generation Track の知見が活かされており、難易度と乱数シードというパラメータを指定することによって、様々な面を生成することが可能となっている。各面には数種類の敵が登場し、敵はそれぞれ独自の動作をしている。エージェントは、これらの敵を避けて進むか、倒して進むかを決定しなければならない。エージェントはマリオを模したものであり、敵にはクリボーやノコノコといったものを模したものが登場する。

本プログラムは、エージェントの周囲の状況 (Fig.1) として、地形情報・敵の位置・種類、地面についているか、ジャンプ可能であるか等を知覚することができ、学習する際の入力情報になる。エージェントは人間が操作するときと同様の行動を取ることが可能で、「左」、「右」、「下」、「ジャンプ」、「加速」の5種類のボタンの組み合わせで行動が表現され、出力情報となる。

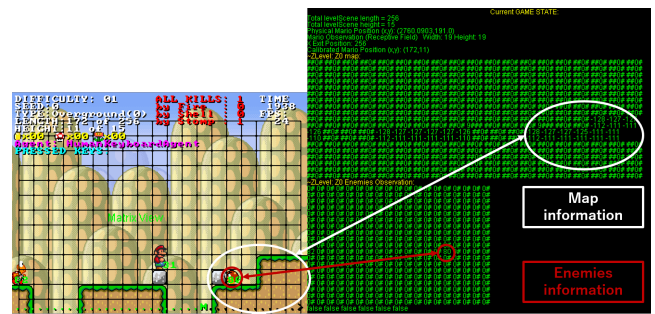


図1 A perceptual information of Mario AI

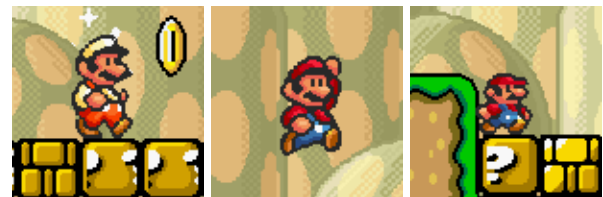


図2 States of agents

エージェントは右方向に進んで行き、制限時間内でゴールに到達する事が目標となっている。エージェントには「ファイアマリオ」「スーパーマリオ」「ちびマリオ」という状態が存在する (Fig.2)。「ファイアマリオ」でダメージを受けた場合「スーパーマリオ」に変化し、同様に「スーパーマリオ」でダメージを受けると「ちびマリオ」に変化する。「ちびマリオ」でダメージを受けた場合は死亡する。穴に落ちた場合はエージェントの状態を問わず、死亡する。

Fig.1 の様な、エージェントを中心とした周囲の状況を生データとして取得する事ができ、学習に用いることができる情報量は 19×19 セル分の範囲である。エージェントは毎フレーム環境情報を取得し、エージェントの行動制御するために毎フレームキー入力を返す必要がある（毎秒24フレーム）。

また、ゲーム内に存在するオブジェクトは全てに対応する値が割り振られており、本研究ではそれらの値をエージェントの学習に用いている。

Mario AI における評価はゲームを1プレイシミュレータに備えられているゲームの評価関数から算出される評価値を得る事である。本研究では、この評価値に基づき各手法の比較及び考察を行った。

表1 Types of competition

Game Play Track	Move the agent once to compete the score
Learning Track	Learning is performed a predetermined number of times using learning method and compete for score
Level Generation Track	Stage creating to compete for score
Turing Test Track	Compete how to make agent behave humanly

3. 関連研究

まず, Mario AI のエージェントの学習には大きく 2 種類の方針がある. 1 つはエージェントを学習させゲーム評価値のみを意識した強いエージェントの追求であり, もう 1 つはエージェントをより人間らしいプレイをするように学習させるものであり, 本研究は前者を取り扱うものとする. 以下, Mario AI 関連の研究としてニューラルネットワークと進化的アルゴリズムを用いた手法について説明する.

3.1 ニューラルネットワーク

ニューラルネットワークは, 脳機能にみられるいくつかの特性を計算機上のシミュレーションによって表現することを目指した数学モデルであり, 一個体の脳を改善し続ける様な学習形態である [6]. ニューラルネットワークは, ある入力に対して線形変換を施し, 活性化関数によって非線形性にも対応する事ができ, 入力に対する何らかの特徴を持った出力を得る事ができる.

Mario AI に対し, ニューラルネットワークの一種である, Deep Learning を強化学習に応用した DQN(Deep Q Network) を用いたものがある [7]. また, 画像ではなく Mario AI で得られる数値データを用いたものとして SRN(Simple Recurrent Network) や MLP(Multilayer perceptron) を用いたアプローチも存在している [8].

ニューラルネットワークの特性として, 入力変数が多くなると予測理由の説明が困難であることから [9], Mario AI においても学習過程がブラックボックス化している.

3.2 進化的アルゴリズム

進化的アルゴリズム (EA:Evolutionary Algorithm) は, 生物界の進化の仕組みを模倣する解探索手法であり, 多数の個体に対し生存競争させるような学習形態である [10]. EA の一種である遺伝的アルゴリズム (GA:Genetic Algorithm) を Mario AI に用いた研究が行われている [5].

Mario AI に GA を用いる場合, 連想配列を用いて画面情報をキーとし, 値に行動を格納する事で遺伝子を生成している. 各画面情報に対して行動の評価値 (ある画面情報に対する行動の優劣を決めるもの) を算出し, その評価値を基に交叉を行う. 次世代に残すかどうかは, 行動の評価値とは別に個体の評価値 (個体の優劣を決めるもの) を用いて選択を行う.

進化的アルゴリズムを用いた場合, SRN や MLP を用いた場合と同様に画面情報を画像ではなく数値データを用いていることが多い [5], [11]. 進化的アルゴリズムは, 多数の探索個体を用いる為評価値計算に要する計算コストの増大を招く為, 画面情報をすべて用いるのではなく情報を圧縮している場合が多い [5].

4. 提案手法

本章では, 提案手法である TLVET (Tabu Learning Visible from Established Tactic) を説明する. まず提案手法の全体について, 次にタブー学習, 最後に学習効率化メカニズムについて述べる.

4.1 フローチャート

本手法では, 解の情報を用いて 1 プレイ行い, 評価値を得ると共に定石が適用できない状況を探し, その状況に対する最適行動の学習を行う.

提案手法の流れを Fig.3 に示すと共に, 以下に各フローにおける処理について概説する.

Step1. 評価

生成された個体群をそれぞれシミュレータの評価関数に通しゲームをプレイする. エージェントが死ぬかタイムアップ, ゴールする事でその個体の評価を終了する.

Step2. 選択

一番評価値の高い個体を次世代に残す. この時母集団サイズ (16) 分だけクローンの生成を行う.

Step3. 判定

高難度状況に陥っているかの判定を行う. ゴール判定と同義であり, ゴールしていなければ Step4. 学習へ遷移し, ゴールしていれば終了する.

Step4. 学習

高難度状況における学習を行う. 高難度状況で 16 種類の行動を試しもっとも評価値の高かった行動を残す. 16 種類の行動は個体番号とリンクしている (Ex. 個体 [1] = 「左」, 個体 [2] = 「右」, 個体 [3] = 「右」 + 「加速」, ...). そして時刻 $T-X$ ($X > 0$) で初期は 3 フレーム分行動を変化させる. そこから一世代経て高難度状況から脱出できなかった場合変化させるフレーム数を +1 する. 一定数変化させるフレーム数を増加させ高難度状況から脱出できない場合, X を +1 し同様の手順を繰り返す.

4.2 タブー学習

Mario AI の問題特性として, 多くの場面において評価値が高くなる特定の行動 (定石) が存在する. 今回は「右」 + 「加速」 + 「ジャンプ」を定石と定義し, 学習していない状況に対してこの行動を行う. 定石を取ったときの結果が良くない場合のみ定石以外の行動を取るように学習する. 本研究では, 下記に示す 3 つの場合を高難度状況と定義した.

- (1) 敵に当たって死んだ場合
- (2) 穴に落ちて死んだ場合
- (3) 一定時間エージェントの位置が更新されない場合

高難度状況に陥った場合, その状況からの脱出を試みる

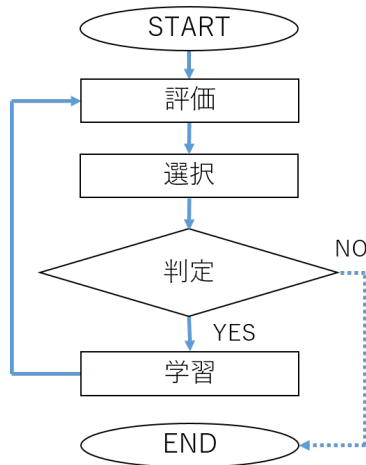


図 3 Flowchart of the proposed method

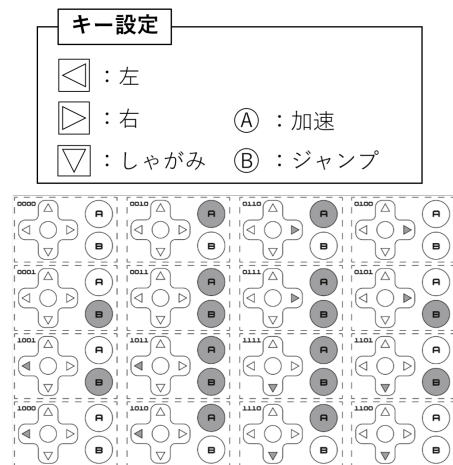
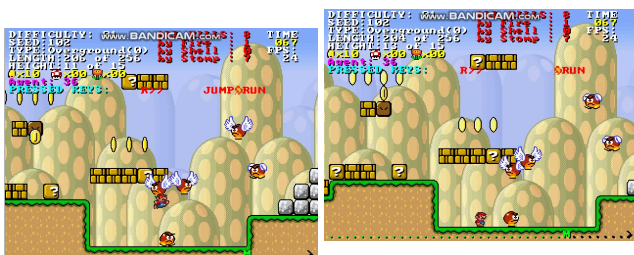


図 5 Details of 16 actions



(a) A high difficulty situation (b) A situation before falling into a high difficulty situation

図 4 About a high difficulty situation

用に学習を行う。定石の行動に「ジャンプ」が含まれている事から、「ジャンプ」後に高難度状況に陥いる事が大半である。そこで、高難度状況 (Fig.4a, 時刻 T) に陥った時、「ジャンプ」直前の状況 (Fig.4b, 時刻 T-1) まで戻りそこから定石以外の行動を取るように学習を行う。

Mario AI は毎フレームエージェントの行動を決めなければならない為、時刻 T-1 という一瞬の時の行動を変化させても高難度状況から脱出出来ないことが多い。その為時刻 T-1 から数フレーム分の行動を変化させる必要がある。

また、時刻 T-1 で一定回数以上フレーム数を変化させて脱出を試みても解決できない場合が存在する。そういった場合時刻 T-1 よりも一つ前の「ジャンプ」した地点に戻る (時刻 T-2)。そして先と同様に数フレーム分の行動を変化させて脱出を試み、脱出できるまで同様の手順で前に戻っていく事で高難度状況の脱出を図る。

つまり本手法における学習とは、“高難度状況を脱出する為にどのタイミングで、どんな行動を何フレーム分行うか”を学習する事になる。

4.3 学習効率化メカニズム

本手法では学習効率化を図る為、行動の制限とランダム性排除及びジャンプ中の空中制御を行っている。以下、これらの仕組みについて説明する。

4.3.1 行動の制限

本手法では、5種類のボタンを組み合わせた全 32 種類の行動には、学習効率化の妨げになる行動が含まれているため、16 種類にまで制限を行った。その 16 種類の行動については Fig.5 に示す通りである。

4.3.2 ランダム性排除

Mario AI での学習ではランダム性を取り入れたものが多く試行毎に結果が異なる。しかし本手法では、高難度状況に陥った場合のみ学習を行い上記にある通り 16 種類の行動しか取らない為そのすべてを試すことでランダム性を排除し、常にその時の最適な行動を選択する事を可能とする。

4.3.3 空中制御

Mario AI は毎フレーム、エージェントの行動を入力する必要がある、その都度入力する行動を何らかの計算で算出しなければならない。エージェントは学習を進めると多くの場合「ジャンプ」を多用する傾向にある。

この「ジャンプ」中にも何らかの行動を入力する必要がある。「ジャンプ」中の入力はその後の結果に影響を与えるが多くの場合でその影響は小さい。そこで TLVET では「ジャンプ」中の入力は「ジャンプ」した時と同じものを入力し続けるものとし、学習量を抑えている。

5. 数値実験

本章では、初めに問題設定について説明した後、本手法と同条件な情報を扱えるニューラルネットワーク手法の SRN(Simple Recurrent Network) と MLP(Multilayer perceptron), 進化的アルゴリズム手法の遺伝的アルゴリズム (GA:Genetic Algorithm) との比較を行い、本手法の有効性を示す。数値実験における比較項目について示す。

実験 1: 定石が通用する面でのスコア比較

実験 2: 定石が通用しない面でのスコア比較

比較 1 では、同評価回数でのスコアを比較する事でコストを膨大に与えた際のポテンシャルを見る事ができ、比較

3 で本手法の基盤であるデフォルト行動が利くか否かでの性能を測り、定石が存在する問題に対するアプローチとしての有効性を示す。

5.1 設定パラメータ

本実験で使用したアルゴリズムの設定パラメータを表 2 に示す。各試行毎に別のシード値を用いている為試行数分の同難易度面での評価を行っている。

5.2 実験 1 (定石が通用する面でのスコア比較) に関する結果と考察

実験 1 では、定石が通用する面でのスコアの伸びを検証する。最終世代のスコアの結果を表 3 に示し、各手法の平均スコアの推移を Fig.6a に示す。

表 3 から分かるように提案手法である TLVET の結果が一番良かった。また、Fig.6a を見ると初期の立ち上がりから最後まで提案手法が常に勝っていることが分かる。このことから他手法と比較し TLVET の探索効率が非常に良い。また、試行毎の平均から見ても分かるように安定して高効率な学習が出来ていると言える。これは面の種類によらず定石が存在し、タブー学習が有効に作用した為だと思われる。よって定石が通用する問題に対して TLVET は有効であることがこの実験から示せた。

5.3 実験 2 (定石が通用しない面でのスコア比較) に関する結果と考察

実験 2 では、定石が通用しない面でのスコア比較をすることで、タブー学習の有効性についての検証を行う。最終世代のスコアの結果を表 4 に示し、各手法の平均スコアの推移を Fig.6b に示す。

表 4 から分かるように提案手法である TLVET の結果が一番良かった。また、Fig.6b から見ても常に他手法と比較して TLVET が優越している事が分かる。しかし、TLVET の最大値が 4096.000 (ゴール) であるのに対し最小値が 1064.852 であり、平均値も 1880.371 とあまり良くない。ゴール出来ている試行が少なく、定石があまり通用していない事が伺える。しかし、定石が通用しない面が多く生成されるであろうパラメータを用いたが、全ての面で定石が通用しないわけではない為、ゴール出来ている試行も存在した。

実験 2 で特に注目したいのが TLVET の最小値の値が他

表 2 Parameters of experiments

Population size	16
Generation size (SRN,MLP)	5000
Generation size (GA)	500
Generation size (TLVET)	100
# Trial	30

手法と比較して大差ない点である。これは、定石が通用しないという結果が顕著に出たものであると考えられ、そういったケースでは他手法と遜色ない結果であると言える。

以上の結果から定石が通用する面では提案手法である TLVET は他手法を圧倒していたが、定石が通用しない面では他手法と圧倒しているとは言えない結果であった。このことから TLVET は想定していた、定石が通用する面に対しては高い学習効率を発揮し、逆に定石が通用しない面に対してはあまり良い結果を示さないというのが確認し、示すことができた。

6. おわりに

本研究では、定石がある問題に対して、効率的なエージェントの学習が出来ることを示した。提案手法は真に学習しなければならない場所のみを容易に作り出しその時のみ学習を行い、無駄な探索を減らすアルゴリズムである。様々な工夫を取り入れているが、すべては探索次元の削減をもたらす為のものである。

数値実験を行った結果、以下の事柄を明らかにする事ができた。

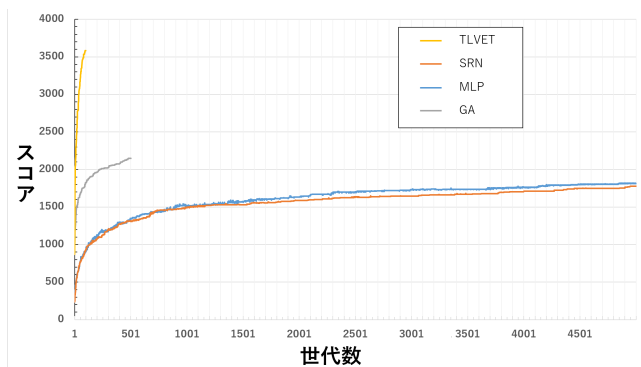
- 提案手法による低次元化の有効性

表 3 Comparison results by each algorithm(Easy case that can be sometimes clear using only standard action)

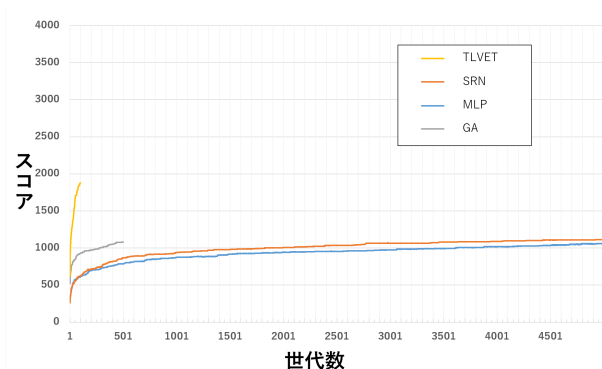
TLVET	Best	4096.000
	Min.	1979.000
	Avg.	3583.445
SRN	Best	2449.815
	Min.	1115.600
	Avg.	1779.139
MLP	Best	2698.731
	Min.	1467.000
	Avg.	1812.279
GA	Best	2989.571
	Min.	1471.310
	Avg.	2150.109

表 4 Comparison results by each algorithm(Difficult case that cannot be clear using only standard action)

TLVET	Best	4096.000
	Min.	1064.852
	Avg.	1880.371
SRN	Best	1582.567
	Min.	835.903
	Avg.	1113.840
MLP	Best	1492.224
	Min.	846.219
	Avg.	1058.533
GA	Best	1432.729
	Min.	751.477
	Avg.	1082.564



(a) Easy case that can be sometimes clear using only standard action



(b) Difficult case that cannot be clear using only standard action

図 6 Comparison of transition of average score by each algorithm

- 他手法と比較した時の探索効率性
- 定石が利く場合に特に有効である結果

比較実験から他手法に対して良い結果を示せたことから提案手法は定石が通用する問題に対し有効である事が分かった。今回の結果から Mario AI だけでなく他問題に対しても定石が存在するものであれば同様のアプローチが可能であると考えられる。

今後の課題として、今回は学習した面に対する評価での比較を行ったが、本手法はエージェントが高難度状況に陥るたびに学習を行うため過学習を起こしにくいと考えている。その為未学習の面に対しても良い結果が得られると推測している。未学習の面での性能向上を図ろうとすると、通常の学習であれば学習すればするほど学んだ面に特化していってしまう。しかし本手法は定石が効かない所のみを学習する為、面に特化することはないと考えており、今後の課題として取り組む予定である。また、本アプローチを他問題に対しての検証も進める予定である。

参考文献

- [1] IEEE Symposium on Computational Intelligence and Games (<http://www.ieee-cig.org/>)
- [2] Sergey, Karakovskiy and Julian, Togelius 「Mario AI Championship site」 (<http://www.marioai.org/>), 2010 年
- [3] Silver, David and Hubert, Thomas and Schrittwieser, Julian and Antonoglou, Ioannis and Lai, Matthew and Guez, Arthur and Lanctot, Marc and Sifre, Laurent and Kumaran, Dharmashan and Graepel, Thore and others 「Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm」 arXiv preprint arXiv:1712.01815, 2017 年
- [4] 半田久志 「Mario AI における Deep Boltzmann Machine を併用した進化学習」 電気学会論文誌 C, 2014 年
- [5] 前田光泰, 蟹井瞳, 中村剛士, 加納政芳, 山田晃嗣 「進化計算を用いたゲームエージェントの行動測の獲得」 Fuzzy

System Symposium, 2012 年

- [6] McCulloch, W.S. and Pitts, W. Bulletin of Mathematical Biophysics (1943) 5: 115. (<https://doi.org/10.1007/BF02478259>)
- [7] KLEIN, Sean. 「CS229 Final Report Deep Q-Learning to Play Mario」.
- [8] Togelius, Julian and Karakovskiy, Sergey and Koutnik, Jan and Schmidhuber, Jurgen. (2009). 「Super Mario evolution」. 156 - 161. 10.1109/CIG.2009.5286481.
- [9] 飯坂達也, 松井哲郎, 福山良和 「構造化ニューラルネットワークの新しい学習法と最大電力需要予測への適用」 電気学会論文誌 B, 2004 年
- [10] Nada M.A. Al-Salami 「Evolutionary Algorithm Definition」 American J. of Engineering and Applied Sciences vol. 2 no. 4 pp. 789-795 2009.
- [11] Intelligent Media Lab - Mario AI manual