

強化学習

Reinforcement Learning

津田塾大学 学芸学部 情報科学科

新田善久

nitta@tsuda.ac.jp

強化学習

- 「対象を支配している法則を知らない状態で試行錯誤を繰り返しながら、望ましい制御方法を学習する」手法
- 何度も試行錯誤しながら、成功した場合は正の報酬を与え、失敗した場合は負の報酬を与えることで、システムは自身の報酬が最大化するように制御ルールを更新していく。
- その結果、各局面でシステムが望ましい行動を取るようになる。

深層強化学習

- 強化学習では、システムに「状態 $s(t)$ 」を与えると、次に行うべき「行動 $a(t)$ 」を出力する。
- ただし、すべての状態におけるすべての「行動」の組み合わせを列挙するには「状態のパターン数」×「行動の種類数」の大きさの表が必要になる。
- この表が大きくなり過ぎると、その問題を扱えなくなってしまう。
- 何らかの方法で「状態を圧縮して表現する」必要がある。
- そのため「深層ニューラルネットワーク」を利用するのが「深層強化学習」である。

強化学習の例: 迷路脱出

- 迷路 縦H×横Wのマス
- マスの意味
 - 水色: スタート地点
 - 赤色: ゴール
 - 白色: 通過できる
 - 灰色: 障害物
- 状態: どのマスにいるか
- 行動: 上下右左のいずれかに移動

nan	nan	nan	nan	nan
nan S0 1.000	1.000 S1 1.000	1.000 S2 1.000	1.000 S3 1.000	1.000 S4 nan
1.000	1.000	1.000	nan	1.000
nan S5 1.000	1.000 S6 1.000	1.000 S7 nan	nan S8 nan	nan S9 nan
1.000	nan	1.000	nan	nan
nan S10 nan	nan S11 nan	nan S12 nan	nan S13 nan	nan S14 nan
1.000	nan	1.000	nan	nan
nan S15 nan	nan S16 nan	nan S17 1.000	1.000 S18 1.000	1.000 S19 nan
nan	nan	nan	nan	nan

強化学習の例：迷路脱出

- ランダムに行動を選択する。

nan	nan	nan	nan	nan
nan S0 1.000	1.000 S1 1.000	1.000 S2 1.000	1.000 S3 1.000	1.000 S4 nan
1.000	1.000	1.000	nan	1.000
1.000	1.000	1.000	nan	1.000
nan S5 1.000	1.000 S6 1.000	1.000 S7 nan	nan S8 nan	nan S9 nan
1.000	nan	1.000	nan	nan
1.000	nan	1.000	nan	nan
nan S10 nan	nan S11 nan	nan S12 nan	nan S13 nan	nan S14 nan
1.000	nan	1.000	nan	nan
1.000	nan	1.000	nan	nan
nan S15 nan	nan S16 nan	nan S17 1.000	1.000 S18 1.000	1.000 S19 nan
nan	nan	nan	nan	nan

強化学習の例：迷路脱出

- 方策 π_0 の確率分布にしたがって、行動を選択する
- 各マスの数字は、その方向に移動する確率

0.000	0.000	0.000	0.000	0.000
0.000 S0 0.500	0.333 S1 0.333	0.333 S2 0.333	0.500 S3 0.500	0.500 S4 0.000
0.500	0.333	0.333	0.000	0.500
0.333	0.333	0.333	0.000	1.000
0.000 S5 0.333	0.333 S6 0.333	0.333 S7 0.000	0.000 S8 0.000	0.000 S9 0.000
0.333	0.000	0.333	0.000	0.000
0.500	0.000	0.500	0.000	0.000
0.000 S10 0.000	0.000 S11 0.000	0.000 S12 0.000	0.000 S13 0.000	0.000 S14 0.000
0.500	0.000	0.500	0.000	0.000
1.000	0.000	0.500	0.000	0.000
0.000 S15 0.000	0.000 S16 0.000	0.000 S17 0.500	0.500 S18 0.500	1.000 S19 0.000
0.000	0.000	0.000	0.000	0.000

強化学習の例：迷路脱出：方策勾配法

- 成功した一連の行動について、その中に現れた (状態, 行動) の重みを更新する。
- 重みにしたかった確率で行動を選択する。

0.000	0.000	0.000	0.000	0.000
0.000 S0 0.500	0.333 S1 0.333	0.333 S2 0.333	0.500 S3 0.500	0.500 S4 0.000
0.500	0.333	0.333	0.000	0.500
0.333	0.333	0.333	0.000	1.000
0.000 S5 0.333	0.333 S6 0.333	0.333 S7 0.000	0.000 S8 0.000	0.000 S9 0.000
0.333	0.000	0.333	0.000	0.000
0.500	0.000	0.500	0.000	0.000
0.000 S10 0.000	0.000 S11 0.000	0.000 S12 0.000	0.000 S13 0.000	0.000 S14 0.000
0.500	0.000	0.500	0.000	0.000
1.000	0.000	0.500	0.000	0.000
0.000 S15 0.000	0.000 S16 0.000	0.000 S17 0.500	0.500 S18 0.500	1.000 S19 0.000
0.000	0.000	0.000	0.000	0.000

強化学習の例：迷路脱出：価値反復法

- 成功したときに、報酬が与えられる。
- 失敗したときは、ペナルティ(負の報酬)が与えられる。
- 時刻 t でもらえる報酬 R_t (即時報酬)
- 未来の報酬を γ ずつ割り引いて考えると
割引報酬和

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

- 行動価値関数 $Q(s,a)$ を表形式で実装する。
- 確率 ϵ でランダムに行動し、確率 $1-\epsilon$ で最大の Q となる行動を選択する (ϵ -greedy 法)

強化学習の例: 迷路脱出: 価値反復法: Q 学習

- Q 学習は Sarsa とは行動価値関数 Q の更新式が異なる。
- Q 学習では、状態 s_{t+1} における行動価値関数の最大値を使用して Q を更新する。

Sarsaの場合:

$$Q(S_t, a_t) = Q(S_t, a_t) + \tau * (R_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$$

Q学習の場合:

$$Q(S_t, a_t) = Q(S_t, a_t) + \tau * (R_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t))$$

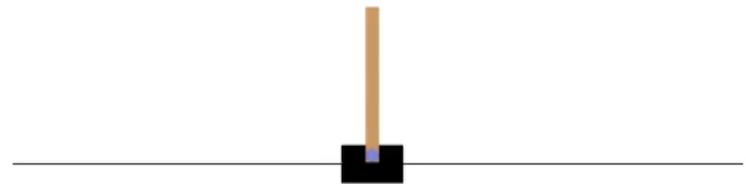
- Q 学習は乱数を使わないので行動価値関数の収束が早い。

強化学習の例：迷路脱出：価値反復法： Q 学習

nan	nan	nan	nan	nan	
nan	S0 0.005	0.071 S1 0.080	0.067 S2 0.052	0.036 S3 0.045	0.033 S4 nan
0.035	0.069	0.070	nan	0.034	
0.069	0.045	0.021	nan	0.009	
nan	S5 0.051	0.091 S6 0.071	0.050 S7 nan	nan S8 nan	nan S9 nan
0.011	nan	0.063	nan	nan	
0.063	nan	0.020	nan	nan	
nan	S10 nan	nan S11 nan	nan S12 nan	nan S13 nan	nan S14 nan
0.086	nan	0.067	nan	nan	
0.077	nan	0.082	nan	nan	
nan	S15 nan	nan S16 nan	nan S17 0.071	0.042 S18 0.092	0.029 S19 nan
nan	nan	nan	nan	nan	

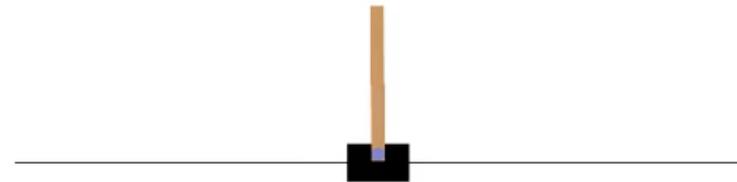
強化学習の例: CartPole(回転振り子課題)

- 台車の上に回転軸を固定した棒を立てて、立てた棒が倒れないように台車を右・左に動かす制御問題。
- 小学生の頃、掃除の時間に、手のひらの上にほうきを立てて倒れないように遊んだ、あのゲーム。
- CartPoleの状態
 - カート位置 $(-2.4, 2.4)$
 - カート速度 $(-\infty, +\infty)$
 - 棒の角度 $(-41.8^\circ, 41.8^\circ)$
 - 棒の角速度 $(-\infty, +\infty)$
- 行動: 右 or 左 にカート进行動かす



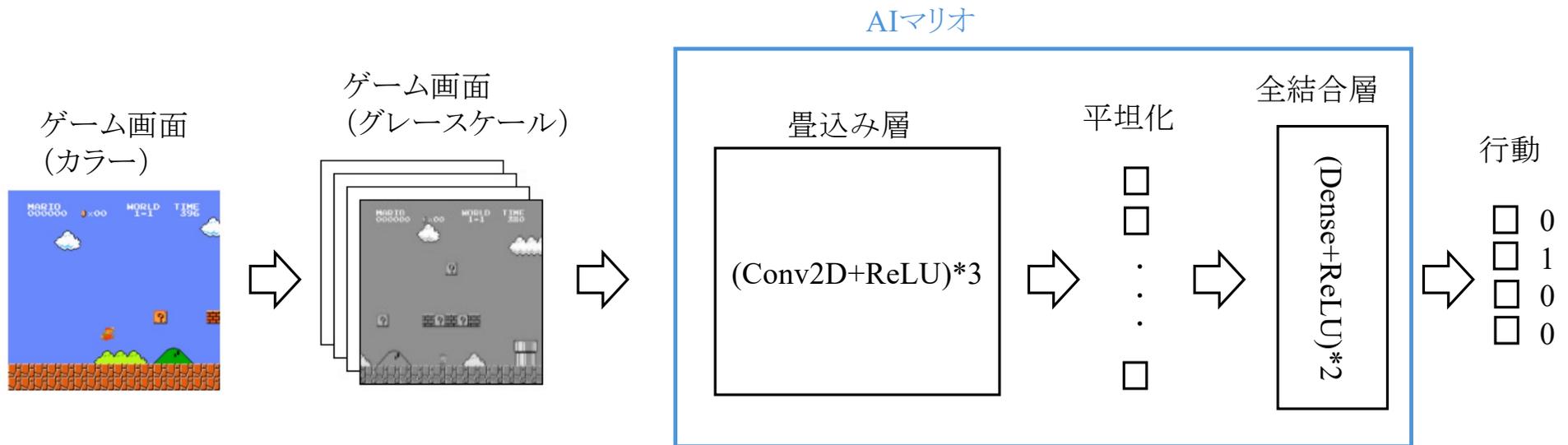
強化学習の例: CartPole: Double DQN

- 倒れないでいれば、報酬を与える
- Main と Target の2つの Q ネットワークを用意する。Target を利用して行動を決定し、学習結果はMain に反映させる。Target は時々 Main で上書きして更新する。(学習が安定する)



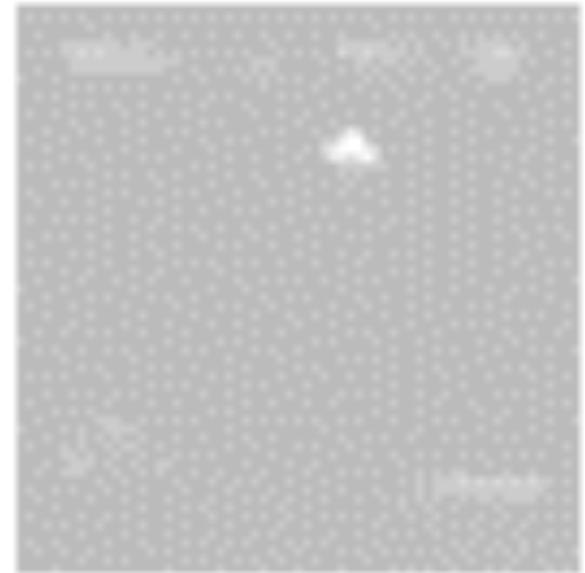
強化学習の例: スーパーマリオ

- ゲーム画面を84x84のグレースケール画像に変換して、4画面まとめてモデルに与え、行動(右,右A,右B,右AB)を出力する。



強化学習の例: スーパーマリオ

- 行動: 右/右ジャンプ/右高速/右高速ジャンプ
- 右に多く進むほど報酬を与える。
- 84x84のグレースケール画像に変換した画面を4画面ごとに与えて、行動を選択させる



強化学習の例: スーパーマリオ: 学習済み

- 40000 エピソードの学習を繰り返したところ、1-1面をクリアできるようになった。

