深層強化学習を用いた株取引エージェントの戦略学習

1　はじめに

近年，機械学習の急速の発展に伴い，深層強化学習を用いた株取引が注目を集めている．特に，深層強化学習の代表的な手法の 1 つに Deep Q-Network(DQN) があるが，それの拡張手法を 6 種類組み合わせた手法である Rainbow を用いた株取引の研究 [1] もなされている．

そこで本実験では，OpenAI Gym の CartPole 問題を Deep Q-Network で実装し, 深層強化学習への理解を深める．次に DQN を用いた株取引エージェントの戦略学習を行うことを目的とする．

2　株取引

東京証券取引所（東証）は日本最大の株式市場である．市場には東証一部，東証二部，マザーズ，JASDAQ の 4 つの区分がある．株式を売買が行われる立会時間は午前 9 時から 11 時半の午前立会 (前場) と午後 12 時半から 15 時の午後立会 (後場) に分けられている．年末年始や休日祝日は休業日である．

2.1　四本値, 出来高

四本値とは設定時間内の始値，終値，高値，安値のことである．始値は設定時間内で初めて取引された株価で，終値は最後に取引された株価である．高値は設定時間内に最も高く取引された株価，安値は最も安く取引された株価である．

出来高とは設定期間内に取引が成立した数量のことである．

2.2　注文方法

株の注文方法には指値注文と成行注文の 2 通りがある．指値注文は自分で取引する値段を決定して注文を行う方法で，必ずしも注文が成立するとは限らない．成行注文は値段を指定しない注文方法である．買い注文であればそのとき出ている最も低い売り注文に対応し，売り注文であれば，そのとき出ている最も高い買い注文に対応するため，即座に注文が成立するという特徴がある．

また，全国の証券取引上において株の注文量は 100 株単位と定められている．

3　要素技術

3.1　OpenAI Gym

OpenAI とは 2015 年に設立された人工知能を研究する非営利企業であり，Gym はその企業が作成した強化学習のシミュレーション用プラットフォームである．

3.1.1　CartPole-v0

CartPole-v0 は，倒立振子を制御する問題である．状態はカートの位置，カートの速度，棒の角度，棒の角速度の 4 変数で表される．カートを左右に動かして棒を安定させ，倒れないようにする方法を学習することが強化学習のエージェントの目標となる．

3.2　Deep Q-Network

Deep Q-Network(DQN) は，強化学習の手法である Q-学習を用いた代表的な深層強化学習の手法である．DQN では深層強化学習に基づく Q-Network と呼ばれる，強化学習における価値に相当する Q 値を多層ニューラルネットにより近似する．Q-Network の更新には状態の遷移を経験として蓄積したものを利用する Exprirnce Replay という工夫がなされている．

3.2.1　強化学習における報酬と価値

強化学習において報酬とは，現在の状態においてエージェントがある行動をとって次の状態に遷移したときに受ける信号である．良い結果を残す行動に正，悪い結果を残す行動に負の報酬を与え，途中の行動に対する報酬は 0 である．

価値は，各状態に対して，それ自体の良し悪しを数値化するために割り当てる．高い価値を持つ状態とは，将来的に期待される利得が高く，特定の方策において高い報酬をもたらす可能性がある状態である．強化学習の手法の 1 つである Q-学習では状態と行動の各ペア

に対する価値を Q 値とし，すべての Q 値は 2 次元の表形式で表現される．

3.2.2 Q-Network

Q-Network とは，Q 値を求める多層ニューラルネットで，入力層には状態変数を入力し，出力層は各行動ごとの Q 値を出力する．連続値の状態空間を扱うことが可能で，一度の入力に基づくニューラルネットの出力計算により全種類の行動の Q 値が得られるため，行動数によって計算量が増えることがほとんどないという利点をもつ．

3.2.3 Experience Replay

Experience Replay とは，過去の遷移情報を保存し，そこからランダムサンプリングすることで，データの時間的相関をなくす工夫である．この経験を蓄積したものを Replay Memory と呼ぶ．

遷移情報は「状態 s_t で行動 a_t を選択したところ，報酬 r_t を獲得し，次の状態が s_{t+1} であった」場合，これらを含む 4 つの組から構成される $(s_t,\ a_t,\ r_t,\ s_{t+1})$ を経験した順に記憶し続ける．設定したメモリの上限を超える場合は最も古い経験から破棄する．

3.2.4 Q-Network の更新

十分に replay memory にデータを蓄えられたら，replay memory からランダムサンプリングし，以下の式に従って Q-Network を更新する．

$$Q_\theta(s_t,a_t) \leftarrow (1-\alpha)Q_\theta(s_t,a_t)+\alpha(r+\gamma \max_{a_{t+1}} Q_\pi(s_{t+1},a_{t+1})) \tag{1}$$

ここで $Q_\theta(s_t,a_t)$ はパラメータ θ を持つニューラルネットワークであり，$Q_\pi(s_t,a_t)$ は教師信号出力用のニューラルネットで，$Q_\theta(s_t,a_t)$ のコピーになっている．$\max_{a_{t+1}} Q_\pi(s_{t+1},a_{t+1})$ は遷移先の状態 s_{t+1} における最大の Q 値，α は学習率 $(0 \le \alpha \le 1)$, r は報酬，γ は割引率 $(0 \le \gamma \le 1)$, t は時刻である．

4 提案手法

本実験では，Deep Q-Network を用いて CartPole 問題を解き，さらに株式取引エージェントの戦略学習を行うことを目的とする．

4.1 CartPole 問題

CartPole 問題では，各ステップの状態を用いてカートを左右どちらに動かすかを Deep Q-Network で学習し，倒立振子を制御させる．

4.1.1 Q-Network, Experience Replay

Q-Network の入力層にはカートの位置 x，カートの速度 v，棒の角度 θ，棒の角速度 ω の 4 つの状態変数を入力し，出力層で左右に動かす行動の Q 値を得る．常に Q 値が大きい行動を選択すると，初めに与えるランダムな値の影響が大きくなるため，ϵ-greedy 法を用いて徐々にランダムな行動から Q 値に従った行動を取るようする．行動をとったあとは Replay Memory に遷移情報を保存し，ランダムサンプリングを行う．後に Q-Network の重みを学習，更新する．教師信号用の $Q_\pi(s_t,a_t)$ は 1 試行が終わるたびに $Q_\theta(s_t,a_t)$ と同じにする．

4.1.2 報酬

報酬は，各ステップで棒が立っていたら 0，倒れたら-1，定めたステップ数以上立っていたら 1 を与える．

4.1.3 学習完了評価

学習完了の基準は各試行の step 数をある試行回数で平均し，それが定めた評価値以上であれば学習完了とする．

4.2 株取引

CartPole 問題で実装した Deep Q-Network を利用して株取引戦略を学習する．ある銘柄の一定期間における四本値と出来高のデータを用いて，DQN エージェントは東証の営業日の取引開始時に 100 株を始値で買うか買わないかの行動を学習する．ただしエージェントが株を取得した場合はその日の取引終了直前にその株を終値で売却するものとする．注文方法は指値注文で，注文は必ず通るものとする．

表 1: CartPole 問題実験設定

最大試行回数	200 回
1 試行の step 数	200 回
学習完了基準となる評価値	195
報酬を与える基準ステップ数	195
学習完了評価の平均計算を行う試行回数	10
割引率	0.99
Experience Replay のメモリ上限	10000

表 2: Q-Network，$Q_\pi(s_t, a_t)$ のネットワークの設定

最適化アルゴリズム	Adam
損失関数	Huber 関数 ($\delta = 1.0$)
学習率	0.0004

4.2.1 Q-Network, Experience Replay

Q-Network の入力層には前日の四本値と出来高の 5 変数を与え，出力層は買うか買わないかの行動の Q 値とする．

Replay Memory は取引終了時の行動が終わったのちに，前日の四本値と出来高，今日の行動，報酬，今日の四本値と出来高を保存する．

4.2.2 報酬設定

ある期間内の学習を 1 試行と定義し，その期間で損益がプラスであれば報酬を 1 与える．1 日ごとの行動では，買って損した場合と買わなかったが株価が上昇した場合は-1 を与え，それ以外は 0 とする． また，試行ごとに異なる期間を学習させて時間的な汎用性を高める．

5 数値実験

5.1 CartPole 問題

表 1 に CartPole 問題における実験設定，表 2,3 に Q-Network および $Q_\pi(s_t, a_t)$ のネットワークの設定，構造を示す．

表 3: Q-Network，$Q_\pi(s_t, a_t)$ のネットワークの構造

層	入力層	隠れ層	隠れ層	出力層
ニューロン数	4	16	16	2

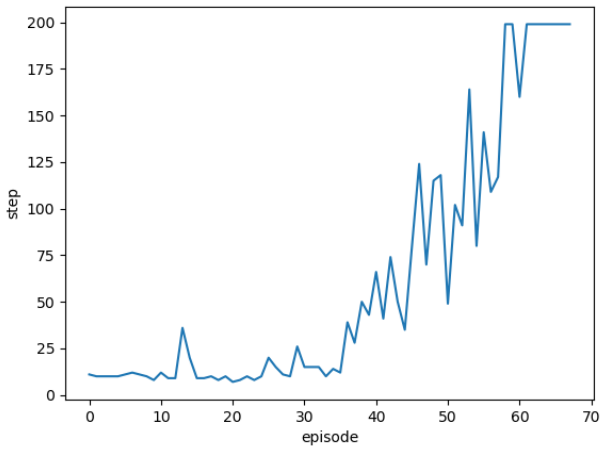

図 1: 試行ごとの step 数の推移 1

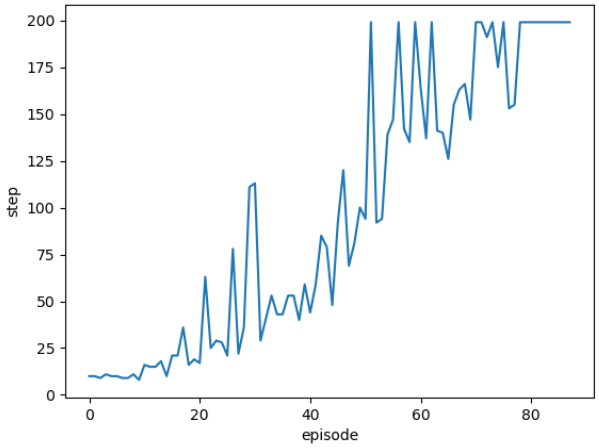

図 2: 試行ごとの step 数の推移 2

5.2 エージェントの戦略学習

扱う銘柄は東証一部に上場している任天堂で，2016 年から 2020 年の 5 年間における四本値と出来高を取得した．また，1 試行は 10 営業日とした．

Q-Network や教師信号用の $Q_\pi(s_t, a_t)$ の実験設定は，ネットワークの入力層を 5，学習率を 0.0005 にし，それ以外は CartPole 問題のときと同じで実験を行った．

6 結果と考察

6.1 CartPole 問題

図 1,2 に結果として得られた試行ごとの step 数の推移の例を示す．

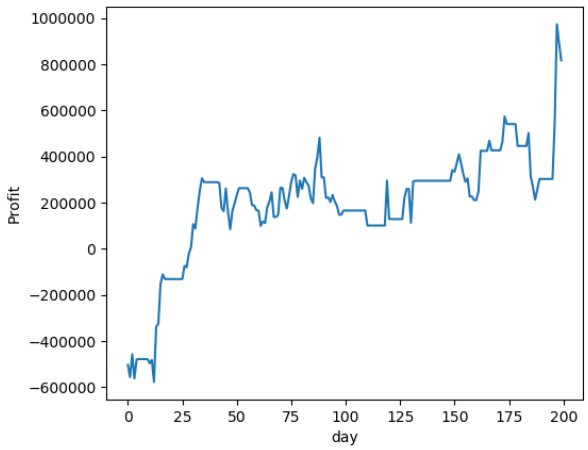

図 3: 直近 200 日の損益の推移 1

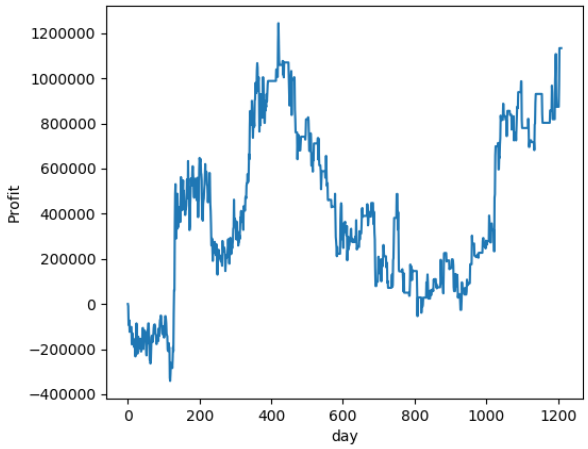

図 4: すべての営業日の取引の損益の推移

どちらも徐々にステップ数が増加し，最後から数試行では 200step 連続で倒立振子を制御することができていることがわかる．何度か実験を行った結果，おおよそ 100 試行あればこれらと同じように学習完了することがわかった．

6.2　エージェントの学習戦略

図 3 に結果として得られた直近 200 日の取引の損益の推移の例，図 4 にすべての営業日の取引の損益の推移の例，図 5 にすべての営業日の任天堂の株価を示す．

エージェントによる学習が進んでいると考えられる直近２００日の損益は増加傾向にあった．しかし図 4 の 400 日目から 800 日目にかけては損益が大きく下がっ

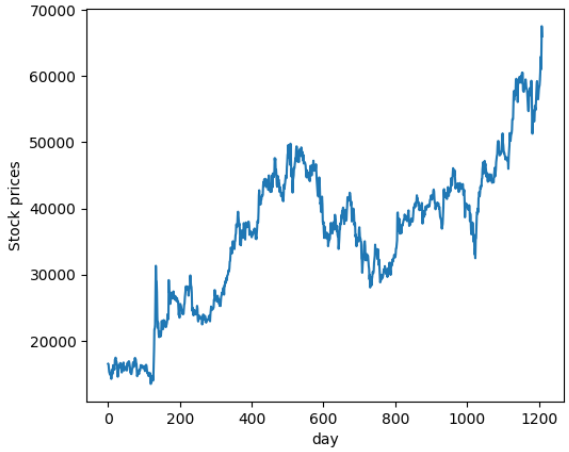

図 5: すべての営業日の任天堂の株価

ており，安定した成果を出すことはできていない．一方，図 5 の任天堂の株価に着目すると，400 日から 500 日頃は上昇傾向で，500 日から 800 日は下降傾向である．このことからエージェントは株価が上昇傾向から下降傾向に変化した後のしばらくの間，株価があがることを期待しすぎているのではないかと考えた．

また，実験を行った 5 年間で buy & hold していた場合約 500 万円の利益を得ることが可能で，今実験でエージェントが得た利益よりも大きかった．

7　おわりに

本実験では，Deep Q-Network を用いて CartPole 問題を解き，株取引戦略を学習した．CartPole 問題は順調に学習が進み，深層強化学習への理解を深めることができた．エージェントの戦略学習については安定した利益をうむ戦略を学習することはできなかった．

今後の課題としては，より発展的な深層強化学習の理解，株取引における学習期間，1 試行あたりに学習する営業日，報酬の与え方などのの調整，さらに板情報や気配値を考慮したより現実的な取引戦略の学習などがあげられる．

参考文献

[1] 森 大典. 深層強化学習 Rainbow を用いたデイトレード戦略の構築. 2018.