Deep Q-Network による Cart-Pole 制御における入力情報の検証

EP13018 稲盛有那　　指導教授：藤吉弘亘

1. はじめに

強化学習は，報酬を最大化する動作を獲得する学習手法であり，深層学習を用いた Deep Q-Network(DQN)[1] が提案されている．DQN は，Q 学習における写像を Deep Neural Network(DNN) で求めることで，画像を入力とした強化学習を実現した．強化学習は未知の環境で学習する場合，学習に時間がかかるという問題がある．本研究では強化学習の高速化を図るために，状態として扱う情報の種類に対する収束性について検証する．

2. Deep Q-Netwark

強化学習は，教師信号の代わりに環境とエージェントのやりとりで得られる報酬を手掛かりに学習する．DQN は，従来の強化学習手法の Q 学習と，DNN を組み合わせた手法である．Q 学習は，ある環境における行動の価値 (以下 Q 値) を，繰り返し更新して獲得する．式 (1) に，Q 値の更新式を示す．

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a} Q(s',a) - Q(s,a)] \quad (1)$$

これより，ある環境 s において行動 a を選択し，報酬 r を獲得する．そして，次の環境 s' で取りうる行動の価値が最大のものを選ぶ．現在の行動の価値と次時間の行動の価値をもとにして，現在の状態における行動の価値を更新する．

DQN では，Q 値の更新を誤差関数と見立てて DNN に用いる．しかし，単純に DNN を適用するだけではパラメータの値が振動や発散することがある．この問題を防ぐために，DQN では Experience Replay，Target Q-Network，報酬のクリッピングを取り入れている．Experience Replay は，試行により獲得した s, a, r, s' をメモリに蓄積する．そして，学習する際にそれらを繰り返しランダムサンプリングして利用する．この処理により，データを与える順番に相関が生じるのを防ぐ．Target Q-Network は，学習の際に一定の更新回数内で NN の重みを固定し，誤差を蓄積して重みの更新を行う．この処理により，学習が収束しやすくなる．報酬のクリッピングは，報酬が正の場合は+1，負の場合は-1 に設定する．この処理により，Q 値の急激な増大を防ぐことができ，勾配が安定する．しかし，報酬の大小を区別できなくなるデメリットがある．

3. 検証内容

本研究では，強化学習の古典的な問題である Cart-Pole 制御を対象とする．Cart-Pole 制御は，図 1 に示すように台車上の棒が直立したまま倒れないように台車を移動させる問題である．台車が取りうる動作は，“右へ移動”か“左へ移動”の 2 つである．この問題において，どのような入力情報を用いることで，強化学習の学習収束性が良いかを検証する．

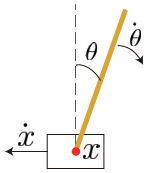

図 1：数値情報として得られる内容

本研究での入力情報の種類を以下に示す．

検証 1: DNN に数値情報を入力

検証 2: DNN に台車周辺画像を入力

検証 3: CNN に台車周辺画像を入力

検証 4: CNN に台車周辺画像+数値情報を入力

図 1 に，数値情報として得られる内容を示す．図 1 の x は台車中心の x 座標，θ は棒の角度，$\dot{x}$ は台車の速度，$\dot{\theta}$ は角速度である．検証 1 の DNN の構造は，入力層のユニット数が 4，中間層は 3 層でユニット数はすべて 100，活性化関数は ReLU を使用する．検証 2 の DNN の構造は，入力層のユニット数が 7056，中間層は 4 層でユニット数は 1000，活性化関数は ReLU を使用する．表 1 に，検証 3,4 の CNN の構造を示す．検証 4 では，画像から得られる棒の角度の視覚情報と，x，$\dot{x}$，$\dot{\theta}$ の数値情報を全結合層に入力する．

表 1：CNN の構造

	処理	詳細
入力層	サイズ	84×84
畳み込み層 1	活性化関数	ReLU
	フィルタ	$3 \times 3 \times 16$
	maxpooling	2×2
畳み込み層 2	活性化関数	ReLU
	フィルタ	$3 \times 3 \times 32$
	maxpooling	2×2
畳み込み層 3	活性化関数	ReLU
	フィルタ	$3 \times 3 \times 64$
	maxpooling	2×2
全結合層	活性化関数	ReLU
	ユニット数	200
出力層	ユニット数	2

4. 検証実験

本実験では，入力する状態に対する学習の収束性として，エピソードに対する獲得報酬を評価する．

4.1. 検証概要

本実験では，OpenAI Gym[2] で提供されている CartPole-v0 の環境を用いる．各ネットワークは，最大エピソード数 1000，1 エピソードにつきエポック数は最大 200，バッチサイズ 100 で学習する．棒の傾きが 15 度未満ならば+1 点，15 度以上ならば 0 点を報酬として与える．棒が 15 度以上傾く，もしくは台車が中心から台車 2.4 個分以上移動するとエピソードを終了する．

4.2. 検証結果

図 2 に，エピソード 0〜129 における獲得報酬を示す．数値情報を入力した場合 (検証 1)，67 エピソード以降は満点である 200 点を獲得した．一方，画像を入力した場合は全て (検証 2-4) 高い報酬を得られず，約 10 点と報酬の最大化に失敗した．以上より，Cart-Pole 制御のようなシンプルな環境においては，数値情報のような行動に直結しやすい情報を入力する方が早く学習が収束することが判明した．

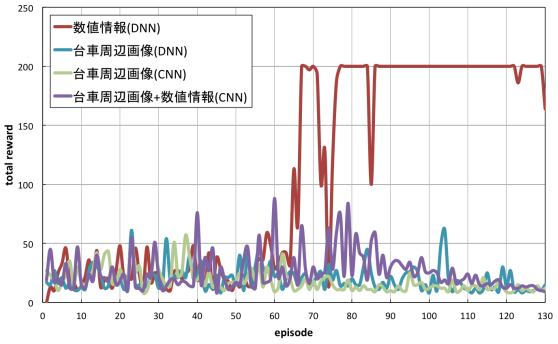

図 2：エピソード毎の獲得報酬

5. おわりに

本研究では，Cart-Pole 制御のようなシンプルな環境において，数値情報を用いて学習する方が有効であることを確認した．今後は，テレビゲームのような複雑な環境での最適な学習について検証する．

参考文献

[1] V. Mnih, *et al.*, “Human-level control through deep reinforcement learning”, Nature, 2015

[2] OpenAI , “OpenAI Gym” , `https://gym.openai.com`