深層強化学習における価値・方策ベースの有効性の傾向調査

ER17073 村上　遥　　指導教授：藤吉弘亘

1. はじめに

深層強化学習は，深層学習と強化学習を組み合わせた学習方法である．強化学習は，ある環境から獲得できる報酬の最大化を目的とし，最適な行動を獲得する学習方法である．深層強化学習手法には，大別して価値ベースと方策ベースがある．深層強化学習は人の性能を超えることもあるが，手法ごとに学習環境や学習条件が異なるため，各手法にどのような傾向があるか明確でない．そこで本研究では，価値ベース・方策ベースで代表的なアルゴリズムの傾向の調査を行う．

2. 深層強化学習

深層強化学習は，深層学習と強化学習を組み合わせた学習方法である．以下に，価値ベースと方策ベースの手法に述べる．

2.1. 価値ベース

価値ベースは，行動の価値の最大化することを目的として，Q 学習を用いて学習する．Q 学習は TD 誤差により，行動の価値を更新する．TD 誤差 δ はある状態 s で行動 a を行うときの行動の価値 $V(s)$ と，行動後の状態 s' における行動の価値 $V(s')$ の差をもとに式 (1) のように算出する．TD 誤差が大きいほど状態 s での行動の価値が高くなる．

$$\delta = r + \gamma V(s') - V(s) \tag{1}$$

ここで，式 (1) の，r は報酬，γ は，割引率である．

Deep Q-Network(DQN)[1]：行動の価値 (Q 値) を畳み込みニューラルネットワークによる近似関数で表現した強化学習手法である．また，Experience Replay や Target Q-Network，Reward Clipping を導入して，膨大な状態数の画像入力に対応している．

Categolical DQN[2]：DQN を基とし，行動価値を分布として表現する強化学習手法である．

2.2. 方策ベース

方策ベースは，行動の規範となる方策を最適化することを目的として，方策勾配法を用いて学習する．方策勾配法は方策 π_θ のパラメータ θ で，収益 (報酬の総和) の期待値を微分して勾配を算出し式 (2) のように更新する．式 (2) の α は学習率を，J は期待収益を示す．

$$\theta^{t+1} \leftarrow \theta^t + \alpha \nabla_\theta J(\theta) \tag{2}$$

Asynchronous Advantage Actor-Critic(A3C)[3]：複数の worker を用いて非同期的にパラメータ更新を行う Asynchronous，2 ステップ以上先の報酬を考慮する Advantage，現在の状態における行動選択と行動を評価を同時に行う Actor−Critic の 3 つを合わせた強化学習手法である．

Trust Region Policy Optimization(TRPO)[4]：更新前と更新後の方策 $\pi_{\theta_{old}}$ との KL ダイバージェンスをしきい値 δ 以下に抑える制約を用いて，方策を改善する強化学習手法である．

Proximal Policy Optimization(PPO)[5]：TRPO を基に更新前と更新後の方策 $\pi_{\theta_{old}}$ の変化量を一定の範囲内にクリッピングことで学習を安定させる強化学習手法である．

3. 価値・方策ベースの有効性調査

複数のタスクにおいて強化学習手法の傾向を調査する．本実験では，使用する強化学習手法として，DQN･Categolical DQN・A3C・TRPO・PPO を用いる．

3.1. 調査環境

環境には，Open AI が提供する Atari 2600 のゲームタスクのうち Breakout，Ms.Pacman，Pong, Qbert, Spaceinvaders を使用する．これらの環境は報酬の種類や，難易度が異なるように選択した．Breakout は，パドルでボールを打ち返しブロックを崩すゲームである．Ms.Pacman は，ゴーストを回避しながらクッキーを食べるゲームである．Pong は，CPU とホッケーを行うゲームである．Qbert は，ジャンプしてブロックの色を変更するゲームである．Spaceinvades は，インベーダを打ち抜くゲームである．各手法は 1.0×10^7 エピソードの学習を行い，1000 回テストした時の平均スコアを算出する．また，ランダムに行動を選択した場合のスコアとも比較する．

3.2. 調査結果

表 1 に各ゲームにおける各手法の平均スコアを示す．表 1 から，Breakout･Qbert･Spaceinvaders では，Categorical DQN，PPO が高スコアを獲得した．Categolical DQN は行動価値を分布で表現することにより，学習に含まれない状態に対する行動価値を正しく求めることができるため，スコアが向上したと考えられる．PPO は方策の変化量のクリッピング効果により学習に含まれる状態にオーバーフィットしないため，スコアが向上したと考えられる．この 3 つのゲームでは，価値ベース・方策ベースのどちらも有効である．

Ms.Pacman では価値ベースの手法の方が高いスコアを獲得した．これは，Ms.Pacman がエピソード後半で獲得できる報酬が少なくなると時，ゲームの進行に合わせて獲得できる報酬が減るため，方策ベースでは適切な行動を得ることができない．Pong は，全ての手法において高スコアを獲得できている．これは，環境のタスクが容易であるため，手法による大きな差が出なかったためであると考えられる．

表 1：1000 エピソード間の平均スコア

		Breakout	Qbert	Spaceinvaders	Ms.Pacman	Pong
	Random	2.2	375.0	75.0	218.4	0.0
価値ベース	DQN	89.2	4325.6	803.5	**4226.5**	19.3
	Categolical DQN	406.5	14099.5	**1483.5**	3995.2	**21.0**
方策ベース	A3C	2.5	4353.2	486.2	1941.6	18.7
	TRPO	2.6	4401.2	876.6	1698.2	**21.0**
	PPO	**427.5**	**15897.6**	1128.5	1680.4	**21.0**

4. おわりに

本研究では，Atari 2600 における様々な強化学習手法の有効性調査を行った．報酬を獲得できない環境には価値ベース，反対に常に一定の報酬を獲得できる環境には Categolical DQN，PPO が有効であることがわかった．この傾向をもとに，既存の強化学習手法を組み合わせたアンサンブル強化学習手法の開発が期待できる．今後は複数の強化学習手法の更なる調査を行う．

参考文献

[1] V. Mnih, *et al.*, “Playing Atari with Deep Reinforcement Learning”, NIPS Deep Learning Workshop, 2013.

[2] M. G.Bellemare, *et al.*, “A Distributional Perspective on Reinforcement Learning”, ICML, 2017.

[3] V. Mnih, *et al.*, “Asynchronous Methods for Deep Reinforcement Learning”, ICML, 2016.

[4] J. Schulman, *et al.*, “Trust Region Policy Optimization”, ICML, 2015.

[5] J. Schulman, *et al.*, “Proximal Policy Optimization Algorithms”, arXiv:1707.06347, 2017.