Mutual DDQN：深層強化学習エージェントの相互学習

EP22126 前田登生　　指導教授：山下隆義

1. はじめに

深層強化学習 (DRL) は，環境との相互作用を通じて累積報酬を最大化する方策を深層ニューラルネットワークで学習する枠組みである．代表的な手法である DQN は，状態 s において行動 a を選択したときの累積報酬の期待値を表す行動価値（Q 値）をニューラルネットワークで近似する．しかし，実際より Q 値を大きく見積もり，学習が不安定化する問題がある．この問題は，Q 値に含まれる誤差（推定誤差）とその誤差が増幅されやすい学習構造に起因する．Double DQN(DDQN) [1] は，この学習構造を改善する代表的な手法であるが，推定誤差を直接抑制する手法ではない．そこで本研究では，相互学習 (DML) を DDQN に導入し，学習の安定化を図る Mutual DDQN を提案する．

2. DDQN

従来の DQN は，選択した行動の Q 値が目標値 y に近づくように学習する．しかし，目標値 y は，次状態 s' における最大の Q 値に依存して求められる構造のため，推定誤差により大きく見積もられた Q 値が目標値 y に反映されやすい．DDQN では，目標値 y を求める際に，行動の選択と Q 値の算出に異なるネットワークを用いる．具体的には，学習によって逐次更新されるオンラインネットワーク Q_θ で次状態 s' において Q 値が最大の行動 a' を選択し，更新を遅らせるターゲットネットワーク Q_{θ^-} で行動 a' の Q 値を算出し，報酬 r と合わせて目標値 y を求める．目標値 y を式 (1) に示す．

$$y = r + \gamma Q_{\theta^-}(s', \mathrm{argmax}_{a'} Q_\theta(s', a')) \tag{1}$$

行動の選択に影響を与えた推定誤差が，Q 値の算出には反映されないため，推定誤差により大きく見積もられた Q 値を目標値 y に用いる傾向を抑制できる．

3. 提案手法：Mutual DDQN

本研究では，異なる初期パラメータを持つ 2 つの独立したネットワーク（$Q_{\theta_1}, Q_{\theta_2}$）が互いの出力を参照しながら学習する DML を DDQN に導入し，推定誤差を直接抑制する Mutual DDQN を提案する．提案手法の概要図を図 1 に示す．ネットワーク Q_{θ_i} の学習には，式 (2) および式 (3) で定義する 2 つの損失関数 $L^{(i)}_{MTD}$ と $L^{(i)}_{KL}$ の和を用いる．ここで $i \in \{1, 2\}$，$j \neq i$ とする．

$$L^{(i)}_{MTD} = \left(r + \gamma\, \mathrm{mean}Q(s', a') - Q_{\theta_i}(s, a)\right)^2 \tag{2}$$

$$L^{(i)}_{KL} = D_{KL}(p_j \| p_i) \tag{3}$$

$L^{(i)}_{MTD}$ は，平均 Q 値を用いた目標値と現在の Q 値の二乗誤差である．ここで Mean TD(MTD) は，DDQN の目標値 y に用いる次状態の Q 値を，2 つのターゲットネットワークで算出した平均 Q 値 $\mathrm{mean}Q(s', a')$ に置き換える手法である．これにより，目標値に含まれる推定誤差を緩和し，学習の安定化を図る．

$L^{(i)}_{KL}$ では，状態 s における 2 つのネットワークの出力に softmax 関数を適用して得られる行動分布 p_1 および p_2 の間で KL ダイバージェンスを最小化する．softmax 関数により行動間の相対関係を分布として表現し，すべての行動で Q 値の相対的な大きさと順位関係を学習する．これにより，ネットワーク間の行動分布を近づけ，行動選択の一致を促進する．

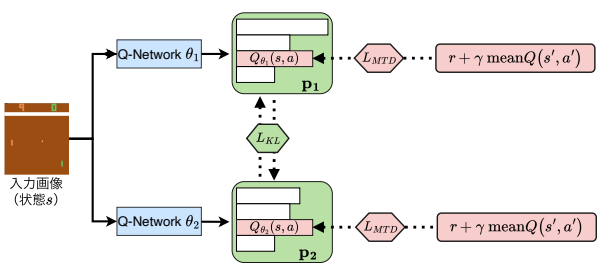

図 1: Mutual DDQN

4. 評価実験

4.1. 実験概要

行動数が異なる複数の Atari 2600 環境のタスク（AirRaid, Enduro, Seaquest）を用い，提案手法の評価を行った．比較手法は，標準的な DDQN および提案手法とする．提案手法に導入した損失関数の分析として MTD のみ，KL のみを使用した比較を行う．学習ステップ数は合計 500 万ステップとする．提案手法の評価には，並列に学習された 2 つのネットワークのうち高い累積報酬を達成した単一ネットワークを用いる．報酬は学習，評価ともに-1 から 1 にクリップする．学習過程は DDQN と提案手法の累積報酬の推移を比較し，学習後の評価指標として，複数エピソードの平均累積報酬，および Q 値と実際の累積報酬の差である Q 値の推定誤差を用いる．

4.2. 実験結果

図 2 に示した累積報酬の推移より，提案手法は学習初期の立ち上がりが早く，全タスクで DDQN を上回る累積報酬の推移を示した．

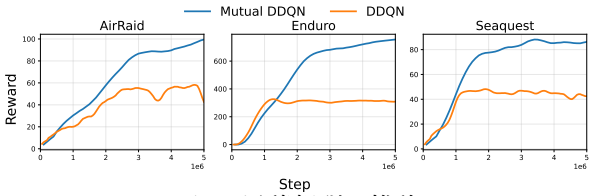

図 2: 累積報酬の推移

表 1 に示すように，全タスクにおいて提案手法は DDQN を上回る平均累積報酬を獲得した．また，MTD のみ，KL のみのいずれも DDQN から性能が向上しており，各損失の導入が有効であることを確認した．

表 1: 平均累積報酬

手法	MTD	KL	AirRaid	Enduro	Seaquest
DDQN			39	323	50
提案手法	✓		88	483	95
		✓	74	834	81
	✓	✓	**119**	**895**	**101**

表 2 に示した Q 値の推定誤差より，全ての条件で推定誤差は正の値を示し，実際の累積報酬よりも推定する Q 値が高くなる傾向を確認した． 提案手法は全タスクで推定誤差が最小であり，Q 値を過剰に見積もる傾向を改善している． MTD のみ，KL のみでも，DDQN と比較して推定誤差は小さいが，KL と比べて MTD がより推定誤差の減少に寄与している．

表 2: Q 値の推定誤差

手法	MTD	KL	AirRaid	Enduro	Seaquest
DDQN			2.1	8.2	2.9
提案手法	✓		0.6	2.5	0.9
		✓	1.0	6.7	2.9
	✓	✓	**0.4**	**2.4**	**0.8**

5. おわりに

本研究では，Q 値推定を改善するために DDQN に DML を導入した Mutual DDQN を提案した．評価実験では，提案手法は従来手法を上回る累積報酬と推定誤差の改善も見られ，提案手法の有効性が示された．今後の展望として，単一ネットワークではなく複数ネットワークの平均 Q 値に基づく行動選択を DML で改善することや Actor-Critic 等の DRL 手法への応用が考えられる．

参考文献

[1] H. van Hasselt et al., *Deep Reinforcement Learning with Double Q-learning*, AAAI, 2016.