自作カードゲーム環境への深層強化学習手法の適用

1 はじめに

近年, 人工知能に関する研究分野は目覚ましい発展を遂げており世の中の様々な分野に応用されている. その中でも人間の脳の働きに近いとされる強化学習と深層学習を組み合わせた深層強化学習は自動運転, ロボット, 推薦システムといった実生活の問題解決だけでなくゲームへの応用も盛んに行われている.

深層強化学習のゲームへの応用は AlphaGo,AlphaZero など将棋や囲碁といった完全情報ゲームへの応用が有名である.AlphaZero は棋譜といった教師データを用いず完全に強化学習のみで既にプロを圧倒している. 一方で最近では麻雀, ポーカーのようなプレイヤーに与えられる情報が部分的である不完全情報ゲームへの応用が注目されている. 本研究では自作した不完全情報ゲーム環境への DeepQNetwork といった深層強化学習の適用, 戦略の構築方法を検証する.

2 要素技術

2.1 OpenAI Gym

OpenAI Gym は人工知能を研究する非営利企業 OpenAI が作った、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. 強化学習の環境として多くのゲームが登録されている. 環境におけるエージェントの行動空間の次元や状態空間, 報酬などを定義することで自作の環境も登録し利用することができる. シミュレーション環境と強化学習アルゴリズム間のインターフェースが確立されているため容易に強化学習を試すことができる.

2.2 Q 学習

強化学習では, エージェントが行動することで環境から報酬を得る. 強化学習における行動はその直後に獲得する報酬の大きさではなく未来に渡っての報酬の総和を見積もった値である「価値」の最大化につながるかという観点で評価される. 価値の最大化を目指す場合にはある状態 s において行動 a をとったときの価値が分かれば良い. この価値のことを Q 値, 行動価値観数と呼ぶ. Q 学習ではエージェントの 1 ステップごとに (1) 式に示す更新式で Q 値の更新を行う.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)) \quad (1)$$

なお, t は時間 , r は報酬, α は Q 値の更新量を現在の Q 値にどれだけ反映させるかを示す学習率, γ は将来の価値をどれだけ割り引いて考えるかを表す割引率である [1].

2.3 Deep Q Network

Q 学習を実際に実装するとなると,Q 値のテーブルができる. しかし, 状態が離散的ではなく連続的な環境の場合 Q テーブルのメモリ量は爆発してしまう. この問題を解決した技術が Deep Q Network である. ニューラルネットワークを用いて, ある状態における行動ごとの Q 値を推定する. エージェントが経験した過去の体験を replay memoly に一定期間保存して置き、過去の経験をランダムにサンプリングして学習を行う Experience Replay, 行動を決定する Q 値のネットワークと Q 値の学習を行うネットワークを分けることで Q 値の過大評価を防ぐ Fixed Target Network といった工夫により安定した学習が可能となっている [2].

2.4 モンテカルロ法

モンテカルロ法は Q 学習と同様に Q 値を推定する学習アルゴリズムであるが, Q 学習のように 1 ステップごとに Q 値を更新するのではなく, 1 エピソードをランダムに行動し終了状態に到達してから Q 値の更新を行う. Q 値の更新は学習率 α , 割引率 γ , エピソードから得られた割引現在価値 G_t を用いて (2) 式に従う [1].

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t)(1 - \alpha) + \alpha * G_t \quad (2)$$
$$where \quad G_t = r_{t+1} + \gamma r_{t+2} + ... + \gamma^{T-t-1} r_{t+k+1}$$

3　作成したカードゲーム環境

未定

4　実験

未定

5　結果

未定

6　今後の課題

未定

参考文献

[1] 久保隆宏. Python で学ぶ強化学習 [改訂第 2 版] 入門から実践まで. 講談社, 2019.

[2] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. Dueling Network Architectures for Deep Reinforcement Learning. *arXiv e-prints*, p. arXiv:1511.06581, November 2015.