深層強化学習に基づく トレーディングカードゲーム環境の構築

第 1 グループ　西村 昭賢

1. はじめに

近年, 人間の学習プロセスから着想を得た強化学習と深層学習を融合した深層強化学習が注目を集めており, 実世界の問題だけでなくゲームにも盛んに応用されている. 本研究ではゲームバランスといった抽象的な概念を深層強化学習と進化型計算という工学的手法によって調整する手法を提案する. また, 独自のトレーディングカードゲーム (TCG) 環境を構築することで数値実験し提案手法の有効性を示す.

2. 要素技術

2.1. Deep Q Network

Deep Q Network (DQN) [1] は代表的な価値ベースの強化学習手法である Q 学習と深層学習を融合させた代表的な深層強化学習手法である.

2.2. Genetic Algorithm

Genetic Algorithm (GA) とは, 生物の進化と進化の過程を模した最適化手法である. GA では多数の個体からなる個体群を用いて解空間の多点を同時に探索するため多目的最適化への応用も可能である.

3. 提案手法

本研究では以下の 3 つの手法を提案する.

1. 数値実験のための独自の TCG 環境.
2. DQN により, 1 の TCG 環境におけるあるデッキの戦略を持つエージェントを構築し, そのエージェント同士で先攻, 後攻でカードを 1 種類ずつ除いて対戦していくことでそのデッキ内の構築戦略下におけるカードパワーを定量的に評価する手法.
3. DQN により得られた結果を基にデッキ内で調整するカードを限定し GA の解空間の次元を削減することで, 調整するカードの枚数を最小限にしながら TCG 環境のゲームバランスを最適化する手法.

4. 実験

4.1. 実験 1

学習側のデッキに恣意的に強いカードを 1 種類, 弱いカードを 1 種類ずつ入れ, 提案する TCG 環境において DQN を適用した. 後攻プレイヤーの行動のみを学習し, 学習後 10000 回対戦を実行した. 学習, 学習後の対戦においては先攻プレイヤーにはルールベースで構築したアグロ, コントロールと呼ばれる戦略を持つエージェントを配置している. アグロは相手プレイヤーに攻撃する戦略で, コントロールは相手盤面のカードの処理を優先する戦略である. また, 先攻プレイヤーは 1 エピソードごとに等確率で戦略を変化させ, 戦略に応じて事前に戦略間の勝率が $50 \pm 5\%$ となるよう調整されたデッキを持つ. 実験から学習済エージェントの勝率, 学習中の獲得報酬の推移を記録し DQN が適用できているかどうか確認した. また対戦した記録から選択した行動, 各カードのプレイされた回数を計測し学習序盤のエージェントと比較することで学習済みエージェントの行動を分析した.

結果として, 0.7182 とベースラインと比較して高い勝率を残すエージェントを構築した. また, エージェントの行動を分析すると盤面のカードで積極的に相手に攻撃して早く相手プレイヤーの HP を 0 にするというアグロに近いがより攻撃的な戦略をとっており, また恣意的に入れた強いカードは学習序盤と比較して優先的に用いられるなどカードの強弱も学習していることを示せた.

表 1: 各手法で得られた最も良好な解の適応度

手法＼適応度	f_p	f_w	f_c
単目的 GA	0.44933	**0.85146**	0.36788
多目的 GA	**0.66365**	0.79097	0.42035
提案手法	0.53259	0.80783	**0.44933**

4.2. 実験 2

実験 1 で構築したエージェントを先攻後攻両方に配置し同じデッキを持たせ, それぞれデッキからカードを 1 種類ずつ除いて 10000 回ゲームを実行して先攻側の勝率を記録した. また比較対象として, アグロの戦略を持つエージェントについても同様に実験をした. 結果として, デッキ内で最も強いと判断できるカードは同一のカードであった一方で, 最も弱いと判断されたカードは異なっていた. これは事前学習の効果により, 学習済エージェント同士の対戦において恣意的に弱く設定したカードは登場回数が少なく勝率計算に及ぼす影響が小さかったためであると考えられる. 以上の点から提案手法により人間のプレイに近い結果を得ることができた.

4.3. 実験 3

TCG 環境に実験 1 から用いてきたデッキをアグロ用のデッキとして追加するといった問題を設定し, 追加するデッキ内のカードのパラメータを環境内のデッキ間の勝率がそれぞれ 50 % に近づくように GA を用いて調整した. デッキ間の勝率の最適化において, 実験 2 のデータから調整するカードの優先順位を付けて提案手法を適用した. また, 比較手法として関連研究 [2] で用いられた勝率を最適化する単目的 GA および勝率とパラメータの変更量を最適化する多目的 GA を適用した. 表 1 に結果を示す. 各解において, パラメータの変更量 p, 勝率 w, 調整するカードの枚数 c についてそれぞれ $f_p = \exp(-p/100)$, $f_w = \exp(-w)$, $f_c = \exp(-c/15)$ と適応度を定義した. 適応度が大きいほどその項目について優れた解である. 各手法において各適応度で優越する解を得ており, 提案手法では調整するカード枚数に関して他の手法より優越した解を得ることができており提案手法の有効性が示された.

5. まとめと今後の課題

本研究では, 深層強化学習を用いた TCG 環境の最適化手法を提案し, 数値実験により有効性を示した. 今後の課題として, DQN とは異なる深層強化学習手法の適用, GA 以外の離散最適化手法の適用, 各アルゴリズムにおける最適なハイパーパラメータの発見などが挙げられる.

参考文献

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. 2013. cite arxiv:1312.5602Comment: NIPS Deep Learning Workshop 2013.

[2] Fernando de Mesentier Silva, Rodrigo Canaan, Scott Lee, Matthew C. Fontaine, Julian Togelius, and Amy K. Hoover. Evolving the hearthstone meta. *CoRR*, Vol. abs/1907.01623, , 2019.