対戦型パズルゲームにおける機械学習 AI と人間の知識を用いた AI の比較

発表者: 総合情報学科 メディア情報学 コース 学籍番号 1310163 柴澤弘樹
指導教員: 橋山智訓 准教授

1 序論

近年のゲーム AI の研究により、AI は人間と同等以上の強さを持つようになってきている。AlphaGo[3] は囲碁で、DQN[2] は Atari 2600 のゲーム 29 種類で、それぞれ人間のプロプレイヤーを上回った。このような強さの背景には、ディープラーニングで評価関数を学習できるようになったことが大きく関わっている。

ディープラーニングでは、従来の人が設計していたものとは異なり、事前知識なしにゲームを学習できた。その汎用性の高さから、様々な問題に応用されている。一方で、学習の結果が分散して記録されるため、その解釈が困難であるデメリットが存在する。内部でどのような処理が行われているかはブラックボックスであり、学習結果が再利用できない。そのため、学習結果の確認やその改善には、多くの計算資源を費やしてトライアル＆エラーを行うしかないのが現状である。

これまでのゲーム AI 研究は強さに着目して行われてきた。すべての情報がプレイヤーに対し開示されている完全情報ゲームの分野では、一定の成果が得られたといえる。しかし情報がプレイヤーから隠されている不完全情報ゲームについては、十分な研究がなされていない。商用ゲームの AI では、開示されていないはずの情報を参照するなど、不公平感を生み出し楽しさを阻害してしまっている。

このような背景から、本研究では不完全情報ゲームにおいて人を楽しませるゲーム AI の実装を目的とする。本稿ではその基礎的検討のために、対戦型パズルゲームである「ぷよぷよ」を対象とし、その対戦 AI の実装を行う。ディープラーニングを用いた AI と人の知識を適用した AI の双方を実装することで、より優れた手法を検討し今後の研究指針とする。

2 ぷよぷよのルール

「ぷよぷよ」は落下型・対戦型パズルゲームの代表作であり、1991 年の発売以後、現在に至るまで広く親しまれている。そのルールを簡単に説明する。

フィールドは横 6 マス縦 13 マスであり、ここに「ぷよ」と呼ばれる色ブロックを設置する。ぷよは 2 つが一組となってフィールドの上部から現れ、自然に落下してゆく。落下中は左右移動および左右回転ができ、任意の列か、隣り合った列にぷよを設置できる。落下予定のぷよは 2 手先まで表示されることから、現在の手を含めた 3 手分の情報が開示されている。一方でそれ以降の手は不明であり、ここに情報の不完全性が存在する。

設置後のぷよの下に空白マスがある場合、ぷよは下へ落下する。同色のぷよを 4 つ繋げると消すことができ、それによって空白が生まれた場合にはその上のぷよが落下する。落下後のぷよが同色で 4 つ以上つながる場合には再び消去が起こり、これが n 回繰り返されることを n 連鎖という。

対戦では連鎖の大きさに応じたスコアが計算される。プレイヤー間のスコアの差に応じて、おじゃまぷよを相手フィールドに降らすことができ、相手の連鎖構築を妨害できる。画面上部の左から 3 列目にぷよを設置するとゲームオーバーとなり、もう一方のプレイヤーが勝利する。

強い AI を実装するためには、連鎖を構築することが不可欠である。素早く大きな連鎖を組むために、不完全な情報から先の手を考慮した十分な先読みと柔軟さを行う事が求められる。今回の実装では連鎖威力、早さ、柔軟さの 3 要素を主眼に置き、これらを満たす AI によって強さを調べる。

3 DQN による戦術の学習

DQN[2] はディープラーニングと Q 学習を組み合わせてゲームプレイを行う AI を構成する手法である。時刻 t におけるゲーム画面を環境 s_t、その時の操作を行動 a_t として、それによって得られた次の時点 $t+1$ における報酬 R_{t+1} から、評価関数 $Q(s_t, a_t)$ を更新する。

評価関数 $Q(s_t, a_t)$ はディープニューラルネットワークによって表現され、そのパラメータ θ_t の更新式は以下の通りである。

$$\theta_{t+1}(s_t, a_t) = \theta_t + \alpha(R_{t+1} + \gamma \max_a Q_t(s_{t+1}, a; \theta_t) - Q_t(s_t, a_t; \theta_t))\nabla_\theta Q_t(s_t, a_t; \theta_t) \quad (1)$$

ここで、α は学習係数、γ は割引係数を表す。

DQN を「ぷよぷよ」のようなパズルゲームに適用した研究は未だなされていない。そこで今回は、スーパーファミコン版の「す～ぱ～ぷよぷよ通　リミックス」を用いて、ゲーム内 AI との対戦を通じた学習を行った。環境には RLE[1] を用い、学習ステップ数 50000 を 100 回繰り返した。報酬 R_{t+1} は、（自スコア - 相手スコア）の時刻 t から $t+1$ での変化とした。

4　人の知識を適用した AI

4.1　ポテンシャル最大化法

従来のぷよぷよ AI において、基礎的な連鎖構築アルゴリズムとしてポテンシャル最大化法が比較対象とされてきた [5, 4]。その手続きは、以下の通りである。

1. 3 手先までの配置可能な手および盤面を全て列挙する。
2. 1 手目でぷよを消去する手を除外する。
3. 2 手目、3 手目で連鎖を発火したとき、スコアが最大となる手を選択する。ただし、候補手が複数ある場合には、その中からランダムに選択する。
4. 選択された手に至る 1 手目を、現在のツモの配置として決定する。

開示されている手に関する全探索を行い、人が定めた評価を用いて最適手を選択する手法となっている。評価方法については、モンテカルロ法を用いた研究 [5] で検討されている。しかし、探索方法と最適手の選択に関する研究は未だ十分なものではなく、改良の余地があると考えられる。そのため、まず 3 手目を開示された手のみに限らず全幅探索とし、またぷよを消去する手も探索に含める改良を行った。

4.2　人の知識の適用

人が評価を定めた探索手法であるポテンシャル最大化法に、さらに人の知識を加えた対戦用 AI の実装を行った。知識は基礎的連鎖である 3-1 階段を構築するための手順を、if–then ルールによって記述したものである。この手順をゲーム開始直後の 6 手分用い、その後に改良した連鎖ポテンシャル法を用いることで、構築連鎖数の安定化を図った。

また対戦のために、構築した連鎖が閾値として定めた威力以上で発動できるとき、即座に発動する戦術をとった。閾値は 4 連鎖相当である、2100 点に設定した。このような単純なルールベース AI を用いて、連鎖構築能力と対戦における強さを調べた。

4.3　連鎖構築シミュレーション

ポテンシャル最大化法の変更に伴う改善効果を検証するため、シミュレーションによって構築連鎖数を調べた。手数は 32 手で連鎖を構築し、33 手目に任意の着手で発火するものとした。試行回数は 50 回とし、すべてのシミュレーションで同じ配石を用いた。

従来のポテンシャル最大化法、探索手法改善による方法、人の知識を適用した手法によるそれぞれの構築連鎖数の分布を、図 1 に示す。探索方法の変更によって連鎖数が大きく改善され、人の知識を適用することでそのさらなる安定化に成功した。人の知識を適用した AI では、配石をより適切に配置し、柔軟で効率的な連鎖を構築できることがわかった。また探索における 1 手あたりの平均計算時間は 53.82 ms であり、十分な早さで大きな威力の連鎖を構築できたと考えられる。

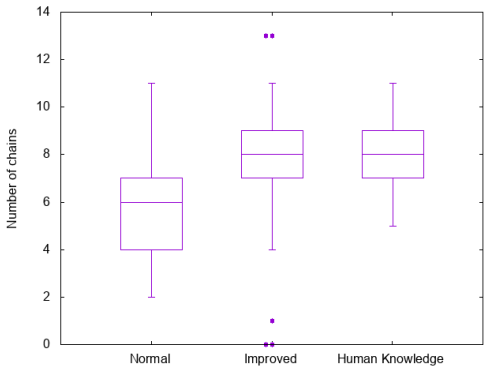

図 1　ポテンシャル法の改良法と人の知識を適用した AI による連鎖構築

5　ゲーム内 AI との対戦

DQN によって対戦の戦術を学習した AI と、人の知識をルールベースで適用した AI について、それぞれゲーム内の AI との対戦成績による強さの比較を行った。対象とするゲームはスーパーファミコンの「す～ぱ～ぷよぷよ通　リミックス」、敵 AI は「のほほ」とした。

それぞれ「のほほ」との 50 戦を行い、その勝敗とスコアを調べた。結果を表 1、表 2 に示す。DQN は時として 5 連鎖を発動するに至るまで学習が進んだが、強さでは圧倒的に劣っていた。一方の人の知識を適用した AI では、単純なルールであったにも関わらず「のほほ」とほぼ対等の強さを誇り、スコアでは上回った。さらに戦術面に関する知識を補強できる余地があり、人の知識を基にする AI はさらに強くなる可能性を秘めている。

表 1　実装 AI と「のほほ」との対戦における勝利数

アルゴリズム	実装 AI	のほほ
DQN	2	48
人の知識を適用した AI	24	26

表 2　実装 AI と「のほほ」との対戦における平均スコア

アルゴリズム	実装 AI	のほほ
DQN	743.30	1790.12
人の知識を適用した AI	4762.52	2703.74

6　結論

本稿では、機械学習による AI と人の知識を適用した AI の双方を実装し、対戦型パズルゲームの「ぷよぷよ」においてその強さを評価した。機械学習の手法として DQN を用いた AI では 50 試合中 2 勝に留まったのに対し、人の知識を適用した AI では 24 勝を収めることができた。よって現状では人の知識を用いた AI が、不完全情報ゲームの一つである「ぷよぷよ」では有効であったと結論づけた。

機械学習による AI は学習方法を変えることで、人の知識を基にした AI では戦術に関する知識を補強することで、さらなる強さを獲得できる可能性がある。しかし、機械学習では学習内容の解釈が困難であるために改良が難しく、人の知識の適用ではその形式化が難しい。今後はこれらを組み合わせることで、より効率的な学習を行い、強さを向上させることを目指す。

参考文献

[1] N. Bhonker, S. Rozenberg, and I. Hubara. Playing SNES in the Retro Learning Environment. *ArXiv e-prints*, November 2016.

[2] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, Vol. 518, No. 7540, pp. 529–533, 02 2015.

[3] David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, Vol. 529, pp. 484–503, 2016.

[4] 富沢大介, 池田心, シモンビエノ. 落下型パズルゲームの定石形配置法とぷよぷよへの適用. 情報処理学会論文誌, Vol. 53, No. 11, pp. 2560–2570, nov 2012.

[5] 大月龍, 前田新一, 石井信. 不完全情報ゲームに対する階層化したモンテカルロ探索とそのぷよぷよへの適用. 電子情報通信学会技術研究報告. NC, ニューロコンピューティング, Vol. 113, No. 500, pp. 275–280, mar 2014.