ぷよぷよの連鎖構成法のための機械学習と人間からの知識抽出に関する基礎的研究

電気通信大学 情報理工学部 総合情報学科　柴澤弘樹, 橋山智訓, 田野俊一

1　序論

近年のゲーム AI の研究の発展により、ゲームにおける 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と呼ばれる機械学習手法の急速な発展がある。

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、学習の結果がネットワーク内に分散して記憶されるため、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。

本研究では、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 の双方を実装し、比較する。両手法の特徴を分析し、今後の研究指針とすることを目的とする。ゲームは対戦型パズルゲームである「ぷよぷよ」を対象とした。

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。連鎖のスコアが高いため、n を大きくする事が勝利につながる。

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、Q 値 $Q(s_t, a_t)$ を更新する。

$Q(s_t, a_t)$ 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 [4, 5]。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。

本研究ではこの手法に、3 手目を全幅探索とし、ぷよを消去する手も探索に含める改良を行った。

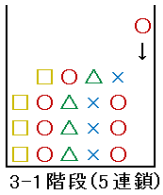

図 1　3–1 階段連鎖の例

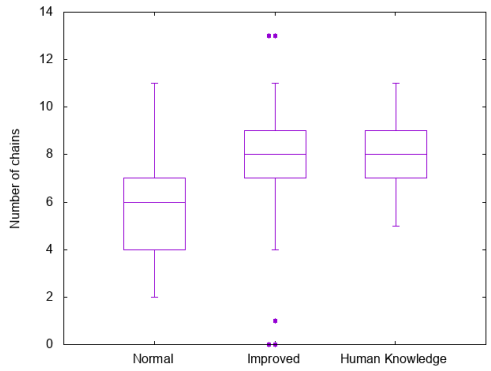

図 2　ポテンシャル法の改良法と人の知識を適用した AI による構築連鎖数

4.2　人の知識の適用

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。知識とは、図 1 に示すような 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て連鎖構築を行った。これにより、安定して大きい連鎖数を構築できることを目指した。

連鎖発動のスコア閾値は、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。

5　実験

5.1　連鎖構築シミュレーション

人の知識を実装した AI に関して、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 2 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より安定的に連鎖可能となった。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、実際のゲームで利用可能な時間であった。

5.2　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。その勝敗とスコアを調べた結果を、表 1、表 2 に示す。人の知識を適用した AI の方が、DQN により得られた AI よりも勝利数、スコアともに成績がよかった。

表 1　実装 AI とゲーム内 AI との対戦における勝利数

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

表 2　実装 AI とゲーム内 AI との対戦における平均スコア

アルゴリズム	実装 AI	ゲーム内 AI
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。人の知識を適用した AI の方が勝利数、スコア共に高かった。人間の知識を適切に記述することは困難な場合もあり、今後機械学習と組み合わせて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, et al. Human-level control through deep reinforcement learning. *Nature*, Vol. 518, No. 7540, pp. 529–533, 02 2015.

[3] David Silver, et al. 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.