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

電気通信大学　情報理工学部　卒業論文発表
総合情報学科　メディア情報学コース
橋山研究室
1310163
柴澤弘樹

背景：ゲームAIの現状

- AIがプロの人間プレイヤーに勝つ
  - AlphaGo……囲碁
  - Deep Q-Learning(DQN)……Atari 2600ゲームの29種類

機械学習（ディープラーニング）の発展

問題

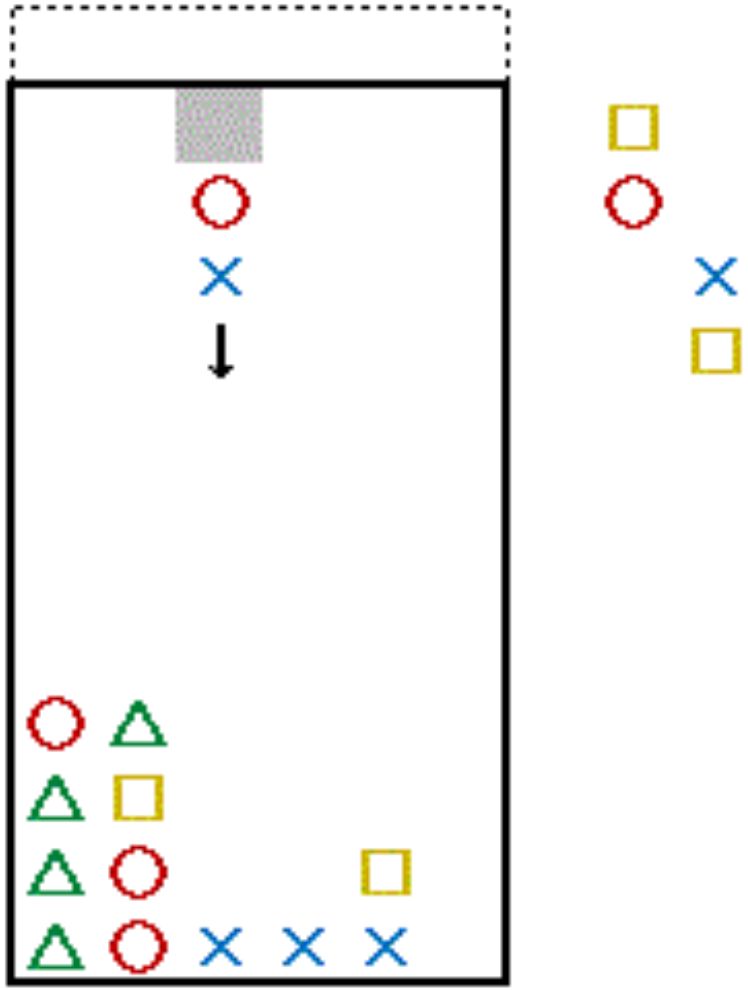
- 内部処理がブラックボックス
- 処理、挙動の解釈が難しい
- 学習結果の再利用が難しい

目的

- 機械学習AI
  - ○ 強い、事前知識不要
  - × 学習結果の解釈が難しい
- ルールベースAI
  - ○ 処理の解釈、改良が容易
  - × 知識のルール化が難しい

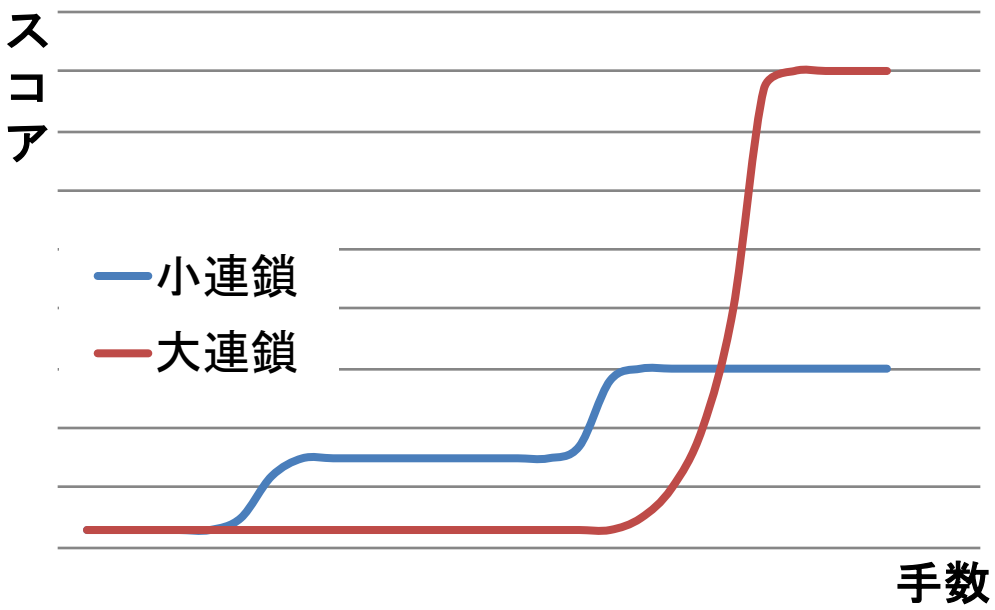
2つの手法を比較、検討

ぷよぷよ



- 連鎖：「ぷよ」を連続で消去
- 連鎖 ➡ スコア ➡ 相手の妨害
- 大連鎖が勝利へ
- 先読み、長期的視点の必要性

連鎖数	最低スコア
1	40
2	360
3	1000
4	2280
5	4840



ポテンシャル最大化法[1][2]

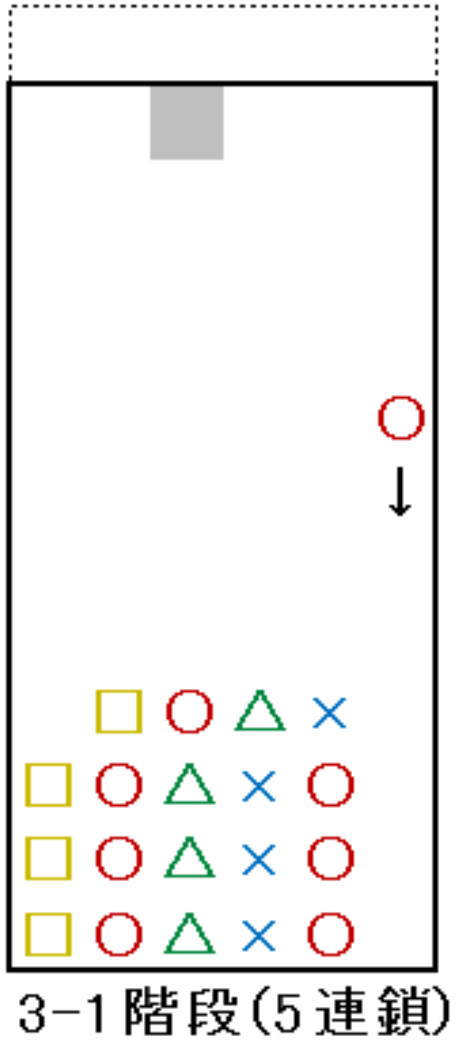
- 見えている手のみ（3手分）を全探索
- 1手目での消去なし
- 2, 3手目で発動する連鎖のスコアを最大化

提案1 → 3手目を全幅探索＋消去の許可

[1] 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.

[2] 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.

人の知識を適用したAI



- 知識
  - 3-1階段の構築ルール
  - if-thenルールを書き下し（設置法20種）

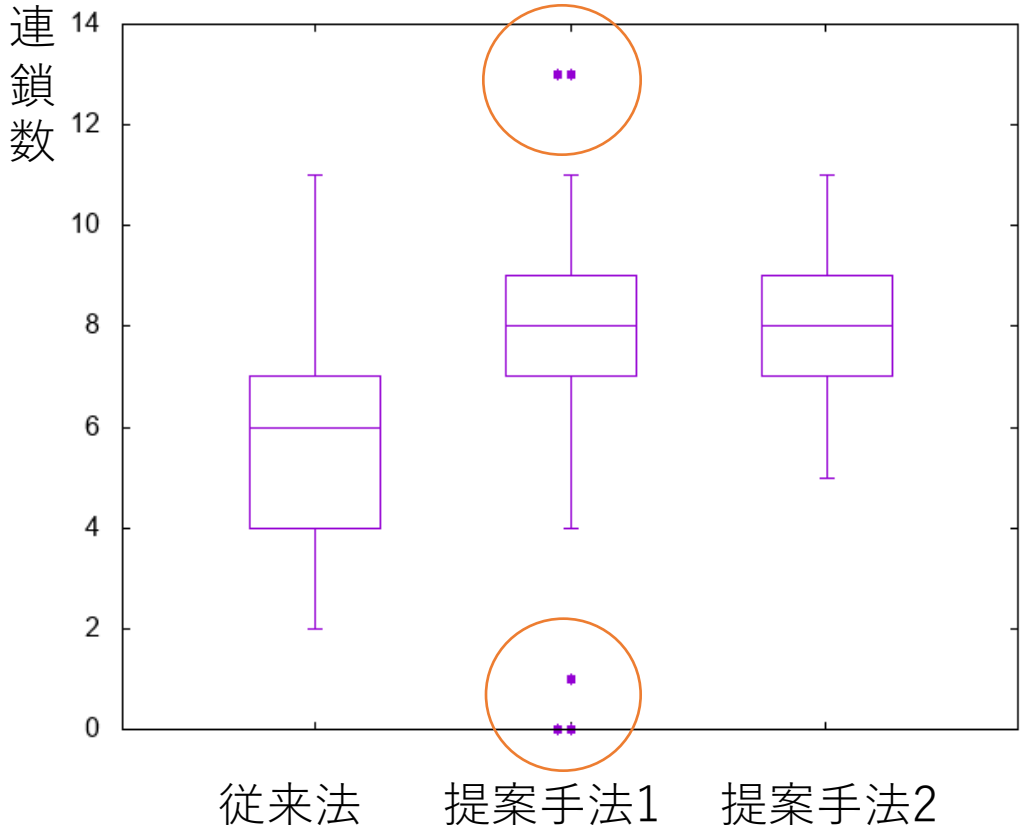
```
            dodai[i+1]++;
            dodaiColor[i+1] = tsumo[0][0];
            return ret;
        }
    }
    for (int i = 1; i < Field.MAX_WIDTH-1; i++) {
        //仕掛け縦
```

提案2 → 初手6手に適用＋提案手法1

- シミュレータを作成

実験1：連鎖構築シミュレーション

- 実験条件：32手+発動1手、50試行（同配石）
- ポテンシャル法、全幅探索、人の知識を適用したAI



平均連鎖数

ポテンシャル最大化法	5.96
ポテンシャル法の改良法(提案1)	7.78
人の知識を適用したAI (提案2)	8.10

人の知識で連鎖数が安定

実験2：間接的な対戦によるAI比較

DQN

- 実装
  DQN-Chainer[3]
  RLE[4]
- 学習
  ゲーム内AIと対戦
  50000ステップ
  ×100回

人の知識を適用したAI

- 画像認識による
  入出力を実装

```
0 0 0 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0
5 4 0 0 0 0
5 4 1 0 0 0
5 4 1 0 0 0
```

- 連鎖発動のスコア
  閾値2100点

ゲーム内AI

- 「のほほ」
- まぐれによる連鎖
- 時として5連鎖以上

[3] https://github.com/ugo-nama-kun/DQN-chainer.git, Last Visited 2017/2/13.
[4] https://github.com/nadavbh12/Retro-Learning-Environment.git, Last visited 2017/2/13.

実験2：対戦結果

勝利数

	実装AI -ゲーム内AI
DQN	2 - 48
人の知識適用AI	24 - 26

平均スコア

	実装AI	ゲーム内AI
DQN	743.30	1790.12
人の知識適用AI	4762.52	2703.74

DQN対ゲーム内AIの様子

人の知識を適用したAIが優れていた

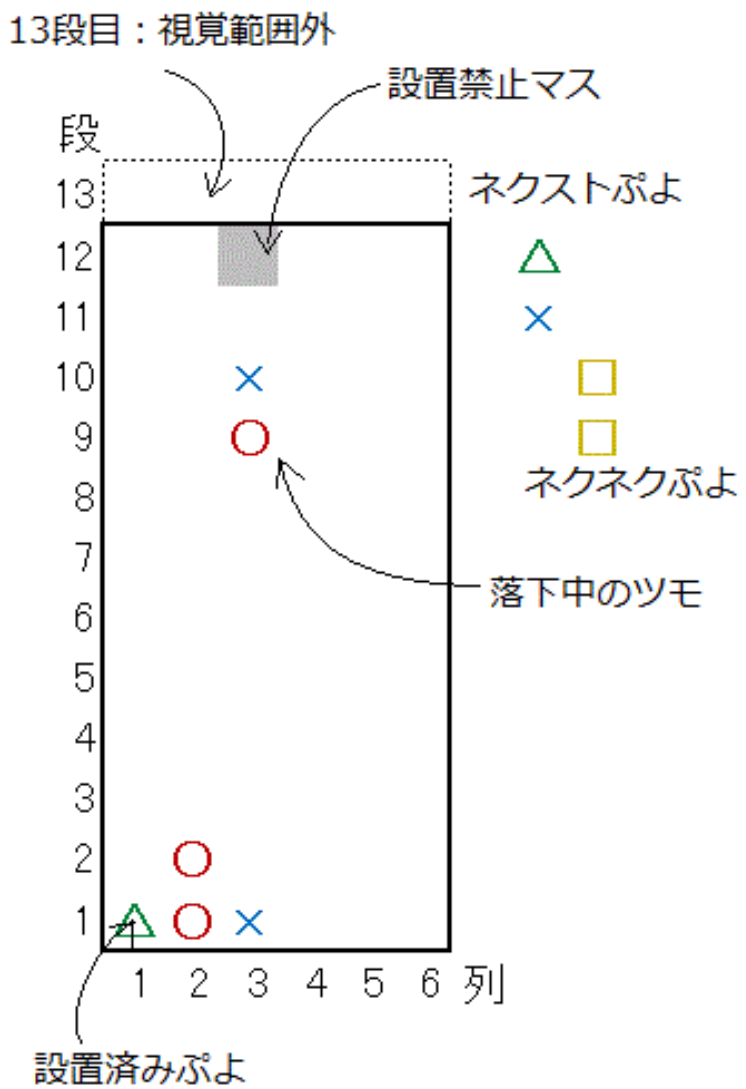
結果のまとめ

人の知識を適用したAIの対戦模様

- 機械学習AI（DQN）
  - ○ 事前知識不要
  - × 弱い、ルールの解釈が困難、改善方針が不明
- 人の知識を適用したAI
  - ○ 強い、ルールの解釈が容易、結果の再利用が可能
  - × 知識のルール化に手間

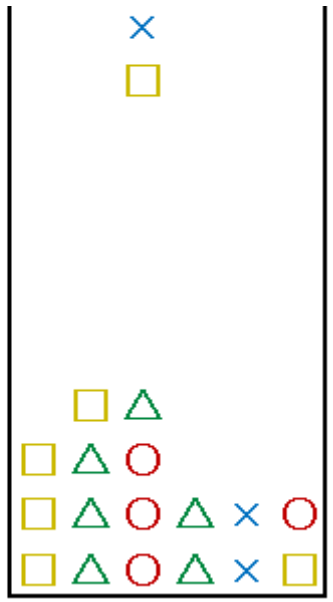
今後　機械学習と知識の組み合わせを検討

補足：ぷよぷよのフィールド

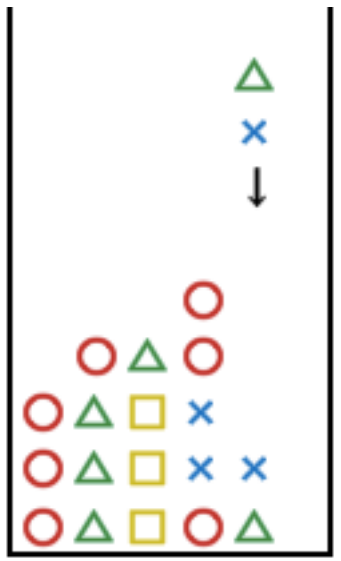


連鎖構築の難しさ

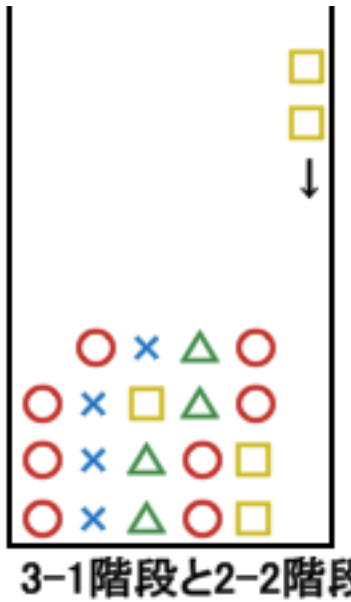
同形異色



置き場なし



途中消し



組み合わせ

DQN: Deep Q-Learning[5]

- Q学習におけるQ値をディープラーニングで学習
- パラメータθ_tの更新式

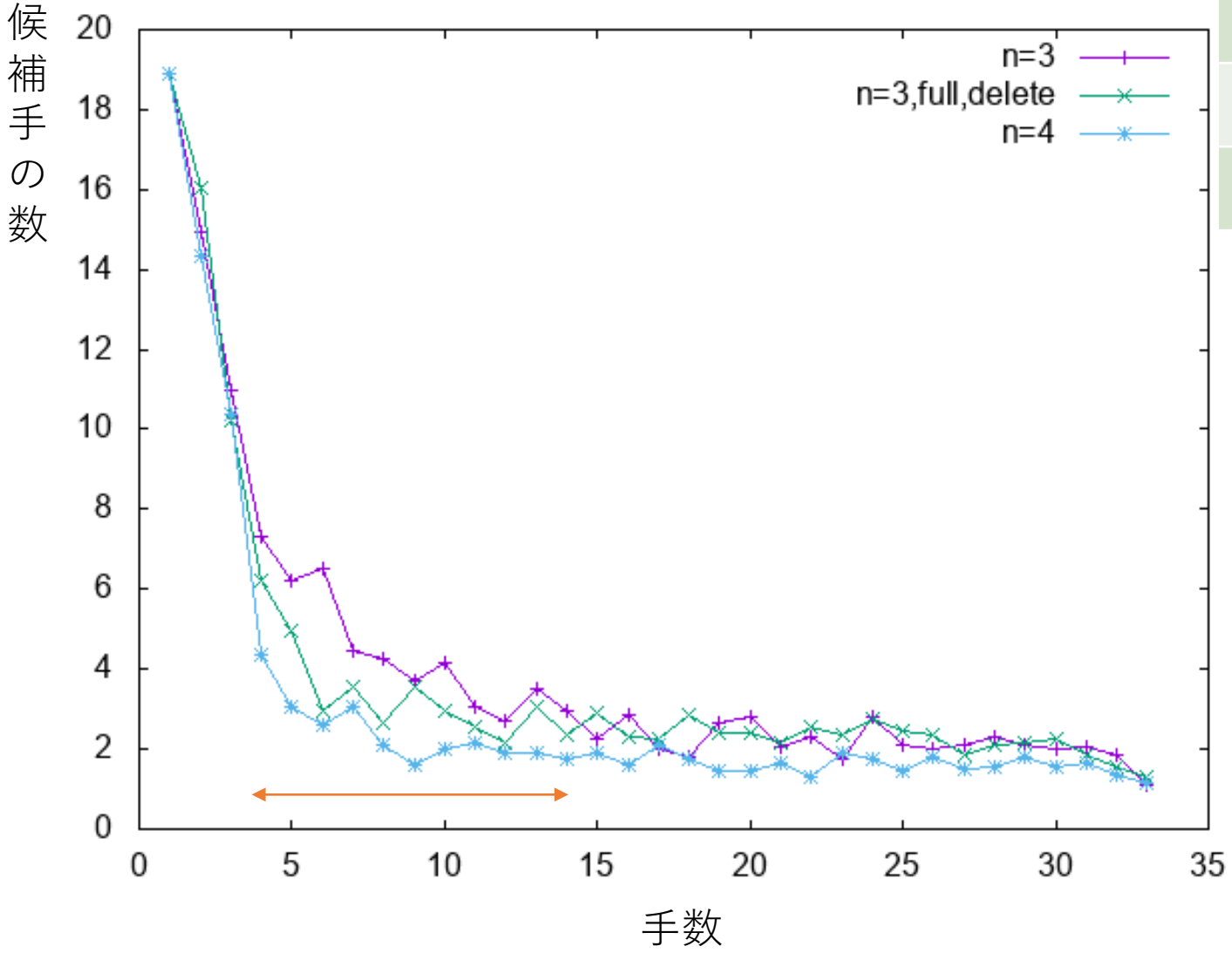
$$\theta_{t+1} = \theta_t + \alpha\left(R_{t+1} + \gamma \max_a Q(s_{t+1}, a; \theta_t) - Q(s_t, a_t; \theta_t)\right)\nabla_{\theta_t} Q(s_t, a_t; \theta_t)$$

t:時刻, α:学習係数, R_{t+1}: $t \to t+1$で得た報酬, γ: 割引係数,
s_t: tでの状態, a_t: tでの行動

- 状態s_t：ゲーム画像4フレーム分
- 報酬R_{t+1}: (自スコア-相手スコア)の変化

[5] Volodymyr Mnih et al. Human-level control through deep rein-forcement learning. *Nature*, Vol. 518, No. 7540, pp. 529–533, 02 2015.

人の知識の適用範囲



平均連鎖数
5.96
8.10
8.36

小

大

- 序盤（4手目から14手目）に差
- 候補手が限られるほど連鎖大？
- 各手数でシミュレーション

➡ 6手が最も良かった