The Hanabi Challenge: A New Frontier for AI Research

Nolan Bard, Jakob N. Foerster, et al.
(Submitted on 1 Feb 2019)

Presented by Joonas Puura
Overview

- Why Hanabi
- How to play Hanabi
- Communication and Theory of Mind
  - The Human approach
- Hanabi Learning environment
- Experiment framework
- Challenges
  - Self-play
  - Ad-hoc teams
- State-of-the-art models
Previously...

- Chess
- Go
- Atari
- Poker (2-player)
- Starcraft*  
  
* not stated in this article
Why Hanabi

- Multi-agent
- Cooperative, non zero-sum game
  - Multiple locally optimal equilibria
  - Efficiency depends on all agents
- Imperfect information
- Communication
Motivation

- Multi-agent interactions are integral part to everything
- AI agents to be capable to interact with other agents / humans
- Consider how other agents act, respond appropriately
- Other agents are typically the most complex part of environment
  - Policies commonly stochastic, dynamic or dependent on hidden information
  - Limited interaction time
Theory of Mind

● “Theory of mind: reasoning about others as agents with their own mental states – such as perspectives, beliefs, and intentions – to explain and predict their behaviour. Alternatively, one can think of theory of mind as the human ability to imagine the world from another person’s point of view.”
● Reasoning about the reason behind the action
Hanabi

- Multi-agent
- Imperfect information
- Cooperative
  - No search for equilibrium
- Solving Hanabi belongs to NEXP complexity class, i.e., requiring exponential time even if P=NP
  - NP-hard even with a centralised, cheating player playing all seats and knowing every hand
Let's play Hanabi, https://www.youtube.com/watch?v=Ofzg71qHh8k
Example of a four player Hanabi game from the point of view of player 0. Player 1 acts after player 0 and so on.
Basic rules - setup

- Player count: 2-5
- Cards
  - 5 colors
  - 5 numbers
  - 5 in hand (2-3 players), 4 in hand (4-5) players
- Deck size
  - 3 Ones, 2 Twos, 2 Threes, 2 Fours, 1 Five
- You see every hand excluding yours
Basic rules 2

- **Score**
  - 0 if lose;
  - Number of cards played otherwise.

- **Hints**
  - Start with 8 tokens
  - Color
  - Rank
  - Empty hints *

- **Losing condition**
  - 3 misplays (3 lives)
  - Try to play, but can’t play
Basic rules 3 - actions

- One action per turn
- Play
  - Misplay - lose health
  - Filled stack - gain information token
- Hint
  - Consumes information token
- Discard
  - Gain additional information token
Simple move

1) P0 hints P1 red
2) P1 plays R1

1 card played with 1 information token
Finesse move

1) P0 hints P2 - Rank 2
2) P1 sees that P2 would misplay if he didn’t do anything
3) P1 plays card from predictable position - plays R1 (topmost card)
4) P2 plays the hinted 2 (R2)

2 cards played with 1 information token.
Communication

- Communication only through legal actions within game
- Explicit - concrete hints - suggesting 4, gives you the information about you having 4 in your hand
- Implicit - conveying extra information implicitly through the choice of actions themselves, which are observable by all players.

- Success of the game depends on efficient communication
  - Sending hints
  - Interpreting hints
What are we learning

- How to give hints
- How to act on given hints
- What and when to discard
- When and what to play ...

- Learning efficient communication convention
Hanabi Learning environment

- Hanabi Learning Environment
- Interface similar to OpenAI Gym
- Provides Rainbow agent

https://github.com/deepmind/hanabi-learning-environment
Challenge One: Self-Play Learning

- Play against yourself
  - 2-5 of the same agent
- Optimise a joint policy to maximize reward in self-play
- Sample limited regime (SL)
  - Sample-efficiency
  - Limit of 100 million interactions - total number of turns
- Unlimited regime (UL)
  - Train model as long as you want
  - Focus on asymptotic performance
Challenge Two: Ad-hoc Teams

- Ultimate goal: playing with other agents or even humans
- Learn to recognize intent in other agents’ actions

- Given ten random self-play games of ad-hoc teammates prior the game
- Reset after each trial (no retaining of games)
- Only two different agents in a game
State of Art

- Experiments using state-of-art-models
- Rule-based agents
- Learning agents
- Self-play vs Ad-hoc play
Rule-Based Approaches

- SmartBot
- HatBot and WTFWThat
- Fireflower
SmartBot

- Rule-based agent
- Tracks publicly known information about each players’ hand
  - Reason about what other players may do
  - What extra information it gains from his point of view
- Plays/discards cards that partner’s might know are safe
  - Therefore other players make assumptions on their hand
- Assumes all players using SmartBot convention
  - Otherwise impossible beliefs
- Parameter specifying uncertain plays
  - Tradeoff between mean score and perfect games
HatBot and WTFWthat

- Makes use of coding theory technique
- Assumes everyone follows convention
- HatBot
  - Predefined protocol to determine recommended action for every other players
  - Reconstruct recommended actions
  - Not very intuitive for humans, but learnable

- Information strategy to directly convey information on each player hands
  - Additional bookkeeping - impractical for humans
  - WTFWthat
Human-style conventions
- Track of private and common know
- Common knowledge on properties of cards (ex. Importance of fives and ones)

2-ply search over all possible actions
- Probabilities on what partner does in response
- Choose the action, which maximizes the value of evaluation function

Evaluation function makes use of physical state of game and belief state
- Ex. commonly known card, output much higher if it is playable

Potentially more natural pairing with human
- Assumes human follows convention

Maximizes win probability
FireFlower conventions 1

- Hints generally indicate play cards, and generally newer cards first.
- Hints “chain” on to other hints, e.g., if A hints to B a playable red 2 as red, then B might infer that it is indeed the red 2, and then hint back to A a red 3 in A’s hand as red, expecting A to believe it is a red 3 to play it after B plays its red 2.
- When discarding, discard provably useless cards, otherwise the oldest “unprotected” card.
- Hints about the oldest unprotected card “protect” that card, with many exceptions where it instead means play.
Hints about garbage cards indicate “protect” cards older than it.

Deliberately discarding known playable cards signals to the partner that they have a copy of that card (with enough convention-based specificity on location that they may play it with absolutely no actual hints).

In many cases, hints about cards that have already been hinted change the belief about those cards in various ways such that the partner will likely change what they do (in accordance with the broader heuristic that one usually should give a hint if and only if they would like to change the partner’s future behavior).
Learning agents

- Actor-Critic-Hanabi-Agent (ACHA)
- Rainbow-Agent
- Bayesian Action Decoder (BAD) agent
Actor-Critic-Hanabi-Agent

- Family of asynchronous advantage actor-critic (A3C) algorithms
- Original is stable, scalable and performant on a range of single-agent tasks
  - Arcade Learning Environment, TORCS driving sim., ..
- Variant successfully applied to multi-agent task of Capture-the-Flag
- Incorporates population-based training for automatic hyperparameter optimization
- Neural architecture:
  - All observations were first processed by an MLP with a single 256-unit 14 hidden layer and ReLU activations, then fed into a 2-layer LSTM with 256 units in each layer.
  - The policy $\pi$ was a softmax readout of the LSTM output, and the baseline was a learned linear readout of the LSTM output.
ACHA - 2 player, param evolution
ACHA - 4 player, param evolution

![Graphs showing score evolution and distribution of points for 4 players.]

- Mean score = 21.57
- Median score = 22.0
- s.d. = 3.68
- n = 1000
ACHA - no param evolution

- 2 player and 4 player, independent runs, local minima problem
ACHA

- Over 20 billion turns

- Parameter evolution hides local minima problem
- ACHA runs with similar final performance can learn different conventions:
  - Color hint to indicate that the 4th card of the teammate can likely be discarded,
  - Color hints to indicate which card of the teammate can be played.
  - Different agents also use the rank-hints to indicate playability of cards in different slot
**Rainbow agent**

- State-of-art agent architecture for deep RL on Arcade Learning Environment
  - Combines key innovations on Deep Q-Networks
  - Sample efficient, achieves high rewards at convergence
- Multi-agent version of Rainbow on Dopamine framework
- Agents controlling different players share parameters
- Feed-forward, no observation stacking outside of the last action

- 2-layer MLP of 512 hidden units each, predict value distributions using distributional reinforcement learning
Rainbow training - 2 agents

![Graphs showing the performance of 2 agents during Rainbow training.]
Rainbow training - 4 agents
Rainbow

- Sample-limited - 100 million turns
- Low variance across independent training runs
- Tend to converge to similar strategies
- Less likely to hint for color compared to ACHA, seemingly no specific convention
- Primarily hints for rank, typically indicating that most recent is playable
- Has one-step memory for past action, no memory for past observations
  - No complex plays?
- Method with high exploration rate - a lot of initial explorations might fail due to running out of lives
Bad-Agent

- Bayesian Action Decoder
- Limited to 2-player self-play setting
- State-of-art for 2-player unlimited regime
- Uses Bayesian belief update
- All agents track a public belief
  - Common knowledge about the cards
  - Posteriors induced from observing actions different agents take
- Explores in space of deterministic policies, while allowing randomness
## Self-play results

<table>
<thead>
<tr>
<th>Regime</th>
<th>Agent</th>
<th>2P</th>
<th>3P</th>
<th>4P</th>
<th>5P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SmartBot</td>
<td>22.99 (0.00)</td>
<td>23.12 (0.00)</td>
<td>22.19 (0.00)</td>
<td>20.25 (0.00)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29.6%</td>
<td>13.8%</td>
<td>2.076%</td>
<td>0.0043%</td>
</tr>
<tr>
<td></td>
<td>FireFlower</td>
<td>22.56 (0.00)</td>
<td>21.05 (0.01)</td>
<td>21.78 (0.01)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>52.6%</td>
<td>40.2%</td>
<td>26.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HatBot</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>22.56 (0.06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14.7%</td>
</tr>
<tr>
<td></td>
<td>WTFWThat</td>
<td>19.45 (0.03)</td>
<td>24.20 (0.01)</td>
<td>24.83 (0.01)</td>
<td>24.89 (0.00)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.28%</td>
<td>49.1%</td>
<td>87.2%</td>
<td>91.5%</td>
</tr>
<tr>
<td>SL</td>
<td>Rainbow</td>
<td>20.64 (0.22)</td>
<td>18.71 (0.20)</td>
<td>18.00 (0.17)</td>
<td>15.26 (0.18)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.5 %</td>
<td>0.2%</td>
<td>0 %</td>
<td>0 %</td>
</tr>
<tr>
<td>UL</td>
<td>ACHA</td>
<td>22.73 (0.12)</td>
<td>20.24 (0.15)</td>
<td>21.57 (0.12)</td>
<td>16.80 (0.13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15.1%</td>
<td>1.1%</td>
<td>2.4%</td>
<td>0%</td>
</tr>
<tr>
<td>UL</td>
<td>BAD</td>
<td>23.92 (0.01)</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>58.6%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Shown are the results for the three learning agents, Rainbow, ACHA and BAD, compared to the rule-based agents, SmartBot, FireFlower and HatBot, for different numbers of players in *self-play*. For each algorithm and number of players we show mean performance of the best agent followed by (standard error of the mean) and percentage of perfect (*i.e.*, 25 point) games. Error of the mean differs based on different number of evaluation games.
### Ad-Hoc results

<table>
<thead>
<tr>
<th>Team ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.84</td>
<td>2.40</td>
<td>0.01</td>
<td>2.03</td>
<td>1.42</td>
<td>1.11</td>
<td>0.10</td>
<td>1.75</td>
<td>1.60</td>
<td>1.32</td>
<td>3.27</td>
</tr>
<tr>
<td>2</td>
<td>2.40</td>
<td>22.12</td>
<td>0.01</td>
<td>1.56</td>
<td>1.05</td>
<td>1.06</td>
<td>0.03</td>
<td>1.10</td>
<td>1.62</td>
<td>0.08</td>
<td>3.06</td>
</tr>
<tr>
<td>3</td>
<td>0.01</td>
<td>0.01</td>
<td>21.67</td>
<td>0.08</td>
<td>0.18</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.05</td>
<td>0.09</td>
<td>2.01</td>
</tr>
<tr>
<td>4</td>
<td>2.03</td>
<td>1.56</td>
<td>0.00</td>
<td>21.72</td>
<td>3.42</td>
<td>1.56</td>
<td>1.13</td>
<td>0.04</td>
<td>1.12</td>
<td>1.02</td>
<td>1.04</td>
</tr>
<tr>
<td>5</td>
<td>1.42</td>
<td>1.69</td>
<td>0.18</td>
<td>3.43</td>
<td>21.24</td>
<td>1.71</td>
<td>1.55</td>
<td>0.02</td>
<td>1.58</td>
<td>1.58</td>
<td>2.03</td>
</tr>
<tr>
<td>6</td>
<td>1.42</td>
<td>1.63</td>
<td>0.00</td>
<td>1.58</td>
<td>1.71</td>
<td>20.03</td>
<td>1.63</td>
<td>0.00</td>
<td>1.24</td>
<td>1.03</td>
<td>1.36</td>
</tr>
<tr>
<td>7</td>
<td>1.13</td>
<td>0.93</td>
<td>0.01</td>
<td>1.13</td>
<td>1.55</td>
<td>1.63</td>
<td>20.74</td>
<td>0.00</td>
<td>1.54</td>
<td>1.22</td>
<td>2.80</td>
</tr>
<tr>
<td>8</td>
<td>0.10</td>
<td>0.02</td>
<td>0.00</td>
<td>0.04</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.91</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.75</td>
<td>1.10</td>
<td>0.05</td>
<td>1.12</td>
<td>1.98</td>
<td>1.24</td>
<td>1.54</td>
<td>0.07</td>
<td>20.98</td>
<td>1.14</td>
<td>1.17</td>
</tr>
<tr>
<td>10</td>
<td>1.60</td>
<td>1.62</td>
<td>0.09</td>
<td>1.82</td>
<td>1.58</td>
<td>1.68</td>
<td>1.22</td>
<td>0.02</td>
<td>1.33</td>
<td>22.48</td>
<td>1.08</td>
</tr>
<tr>
<td>R</td>
<td>1.32</td>
<td>0.98</td>
<td>0.01</td>
<td>1.84</td>
<td>2.03</td>
<td>1.38</td>
<td>0.92</td>
<td>0.00</td>
<td>1.17</td>
<td>1.88</td>
<td>29.52</td>
</tr>
</tbody>
</table>

Score range: 0.0 to 25.0
Results

- Learning agent BAD performed well in 2P;
- In 3-5P rule-based approaches were better than agents;
- None of the agents explicitly trained for self-play succeeded in ad-hoc setup.
Key points

- Hanabi is a multi-agent challenge with combination of cooperative gameplay and imperfect information
- Theory of mind appears to be important
  - Efficient and interpretable communication
  - Capability to reason with humans
- Hanabi Learning Environment
- Currently rule-based approaches (hard-coded bots) are better
- Techniques for self-play fail for ad-hoc
Conditional action probabilities
| $P_{\pi}(a_{t+1}|a_t)$ | Discard  | Play | Hint Colour | Hint Rank |
|------------------------|--------|------|-------------|-----------|
|                        | 1 2 3 4 5 | 1 2 3 4 5 | R Y G W B | 1 2 3 4 5 |
| $a_t$                  |        |      |             |           |
| Discard                |        |      |             |           |
| 1                      | 3 4 7 3 6 | 4 3 1 0 1 | 6 6 5 6 6 | 11 7 7 5 9 |
| 2                      | 4 5 7 2 6 | 4 2 2 0 0 | 6 6 5 6 5 | 12 6 8 5 9 |
| 3                      | 4 4 9 2 7 | 5 3 1 0 0 | 5 6 5 6 5 | 10 6 9 5 8 |
| 4                      | 4 4 7 3 12| 4 3 1 0 0 | 4 5 3 4 6 | 11 8 6 5 9 |
| 5                      | 1 2 4 1 9 | 4 3 1 42 0| 2 2 2 2 2 | 3 6 7 5 2 |
| Play                   |        |      |             |           |
| 1                      | 3 3 5 3 6 | 9 5 2 0 2 | 4 3 4 5 4 | 8 8 7 6 11|
| 2                      | 4 4 6 4 7 | 8 5 1 0 1 | 4 4 4 5 3 | 9 7 8 5 11|
| 3                      | 3 3 4 2 7 | 5 3 1 0 1 | 5 4 4 4 4 | 12 15 8 4 9|
| 4                      | 4 4 5 4 9 | 7 4 1 0 1 | 4 5 4 5 5 | 9 9 9 5 5 |
| 5                      | 5 5 6 5 10| 13 7 1 0 0| 3 3 3 3 3 | 6 7 6 4 5 |
| Hint Colour            |        |      |             |           |
| R                      | 4 5 3 38| 16 14 1 0 0| 2 1 1 1 2 | 3 3 2 2 3 |
| Y                      | 4 3 3 45| 15 12 1 0 0| 1 2 1 1 1 | 3 3 2 1 2 |
| G                      | 5 4 2 37| 19 14 1 0 0| 1 1 2 1 2 | 3 3 2 1 2 |
| W                      | 3 3 3 44| 15 13 1 0 0| 2 1 1 1 1 | 3 2 2 1 2 |
| B                      | 3 5 1 44| 15 12 1 0 0| 1 1 1 1 2 | 3 2 2 2 2 |
| Hint Rank              |        |      |             |           |
| 1                      | 0 0 0 0| 9 6 9 12 64| 0 0 0 0 0 | 0 0 0 0 0 |
| 2                      | 3 4 2 10| 2 4 4 22 42| 0 0 0 1 0 | 8 2 1 0 1 |
| 3                      | 2 3 1 16| 5 6 3 26 24| 1 1 1 0 1 | 6 3 1 0 1 |
| 4                      | 1 1 0 3| 2 2 2 55 31| 0 0 0 0 0 | 0 0 0 0 1 |
| 5                      | 2 2 2 18| 6 3 1 55 4| 1 0 1 1 1 | 0 1 1 1 1 |

Table A.5: Conditional action probabilities for first two player ACHA agent.
| $P_{%}(a_{t+1}|a_t)$ | Discard | Play | Hint Colour | Hint Rank |
|----------------------|---------|------|-------------|-----------|
| $a_t$                | 1       | 2    | 3           | 4         | 5         | 1       | 2       | 3       | 4       | 5         |
| Discard             | 1       | 2    | 3           | 4         | 5         |         |         |         |         |           | 12      | 14      | 11      | 12      | 11       |
|                     | 2       | 6    | 5           | 2         | 12        |         |         |         |         |           |         | 14      | 18      | 13      | 11      | 5         |
|                     | 3       | 2    | 6           | 11        | 2         | 13       |         |         |         |           |         | 18      | 17      | 11      | 4       | 1         |
|                     | 4       | 1    | 2           | 5         | 4         | 19       |         |         |         |           |         | 18      | 14      | 10      | 6       | 2         |
|                     | 5       | 1    | 1           | 1         | 0         | 30       |         |         |         |           |         | 10      | 12      | 16      | 13      | 6         |
| Play                | 1       | 4    | 5           | 3         | 1         | 18       |         |         |         |           | 13      | 10      | 10      | 11      | 9         |
|                     | 2       | 4    | 6           | 5         | 1         | 22       |         |         |         |           |         | 11      | 12      | 11      | 11      | 6         |
|                     | 3       | 4    | 5           | 2         | 1         | 26       |         |         |         |           |         | 7       | 6       | 8       | 14      | 10        |
|                     | 4       | 3    | 3           | 4         | 0         | 0        |         |         |         |           |         | 49      | 10      | 12      | 3       | 3         |
|                     | 5       | 2    | 4           | 5         | 1         | 20       |         |         |         |           |         | 12      | 14      | 12      | 10      | 5         |
| Hint Colour         | R       | 6    | 31          | 25        | 9         | 3        |         |         |         |           | 9       | 4       | 2       | 5       | 2         |
|                     | Y       | 19   | 31          | 12        | 5         | 8        |         |         |         |           | 0       | 0       | 0       | 0       | 1         |
|                     | G       | 20   | 30          | 18        | 1         | 9        |         |         |         |           | 2       | 0       | 1       | 0       | 0         |
|                     | W       | 18   | 18          | 16        | 10        | 15       |         |         |         |           | 0       | 0       | 0       | 0       | 0         |
|                     | B       | 12   | 8           | 11         | 6         | 37       |         |         |         |           | 2       | 0       | 0       | 0       | 0         |
| Hint Rank           | 1       | 3    | 3           | 6         | 8         | 1        |         |         |         |           | 3       | 0       | 1       | 0       | 0         |
|                     | 2       | 3    | 6           | 1         | 0         | 0        |         |         |         |           | 3       | 1       | 0       | 0       | 0         |
|                     | 3       | 1    | 5           | 2         | 0         | 1        |         |         |         |           | 0       | 1       | 0       | 0       | 0         |
|                     | 4       | 1    | 4           | 3         | 1         | 1        |         |         |         |           | 6       | 2       | 1       | 1       | 0         |
|                     | 5       | 2    | 4           | 0         | 0         | 2        |         |         |         |           | 0       | 1       | 0       | 0       | 0         |
| $P_{%}(a_t)$        | 3       | 5    | 4           | 2         | 14        | 5        | 7        | 2        | 1        | 17       | 0       | 2       | 1       | 1       | 1         |

Table A.2: Conditional action probabilities for a first two player Rainbow agent.
Sources

- https://www.reddit.com/r/MachineLearning/comments/anjeal/the_hanabi_challenge_a_new_frontier_for_ai/
- https://en.boardgamearena.com/#lgamelobby