TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Deep reinforcement learning for the cooperative card game Hanabi

Master thesis (sep 2020 ~ apr 2021)

Bram Grooten

Department of Mathematics and Computer Science

Contents

1. How to play Hanabi
2. My project
3. Frontier for AI

How to play Hanabi

2-5 Players

Cards:

- rank (1,2,3,4,5)
- color (B,R,W,G,Y)

Build fireworks

Can’t see your own cards!

How to play Hanabi

One option per turn:

- Give a hint
- Discard a card
- Play a card

Game ends if:

deck empty, 3 lives lost, perfect score

My project

1. Describe multiple DRL algorithms thoroughly
2. Program and analyze a working DRL algorithm
3. Analyze game-theoretical properties of Hanabi

My project

1. Describe multiple DRL algorithms thoroughly
   - Actor-Critic
   - DQN

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha \cdot \delta_t \cdot \nabla_{\boldsymbol{w}_t} V(\tau_t)$$
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \beta \cdot \delta_t \cdot \nabla_{\boldsymbol{\theta}_t} \log \pi(\boldsymbol{a}_t | \tau_t)$$

Notice: action-observation history τ, instead of state $\boldsymbol{s}$.

My project

2. Program a working DRL algorithm
   - Hanabi Learning Environment
   - Linux dual boot

```
TURN OF AGENT NR:  2
Life tokens: 3
Info tokens: 0
Fireworks: R5 Y1 G2 W4 B2
Hands:
Cur player
XX || R1|R1
XX || XX|RGW34
XX || X5|RGW5
XX || XX|RYGW1234
XX || XX|RYGWB1234
-----
G4 || G4|G4
B3 || X3|YB3
Y3 || XX|RYGWB12345
G1 || XX|RYGWB12345
Y1 || XX|RYGWB12345
-----
Y5 || X5|YGWB5
R2 || R2|R2
W1 || XX|RYGWB12345
G2 || XX|RYGWB12345
Y4 || XX|RYGWB12345
Deck size: 3
Discards: W4 B3 G4 Y2 R3 G3 W1 Y2 G5 Y4 B1 R4 B5 B4 Y1 Y3 B1 G1
```

My project

3. Analyze game-theoretical properties of Hanabi

- # states $>$ # deck configurations $= \frac{50!}{(3!)^5(2!)^{15}} \approx 10^{56}$
- small fraction of games is not playable to 25 points
- maximum game length: 89 turns

Frontier for AI

- Multi-agent reinforcement learning
- Cooperative instead of adversarial
- Partially observable
- Hopefully: understand intentions and beliefs of others (humans)

Artificial Intelligence 280 (2020) 103216

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

The Hanabi challenge: A new frontier for AI research

Nolan Bard [a,*,1], Jakob N. Foerster [b,1,2], Sarath Chandar [c], Neil Burch [a], Marc Lanctot [a], H. Francis Song [d], Emilio Parisotto [e,3], Vincent Dumoulin [c], Subhodeep Moitra [f], Edward Hughes [g], Iain Dunning [d], Shibl Mourad [f], Hugo Larochelle [c], Marc G. Bellemare [c], Michael Bowling [a]

[a] DeepMind, Edmonton, Canada
[b] University of Oxford, Oxford, United Kingdom of Great Britain and Northern Ireland
[c] Google Brain, Montreal, Canada
[d] DeepMind, London, United Kingdom of Great Britain and Northern Ireland
[e] Carnegie Mellon University, Pittsburgh, United States of America
[f] DeepMind, Montreal, Canada

ARTICLE INFO

Article history:
Received 10 January 2019
Received in revised form 30 October 2019
Accepted 18 November 2019
Available online 27 November 2019

Keywords:
Multi-agent learning
Challenge paper
Reinforcement learning
Games
Theory of mind
Communication
Imperfect information
Cooperative

ABSTRACT

From the early days of computing, games have been important testbeds for studying how well machines can do sophisticated decision making. In recent years, machine learning has made dramatic advances with artificial agents reaching superhuman performance in challenge domains like Go, Atari, and some variants of poker. As with their predecessors of chess, checkers, and backgammon, these game domains have driven research by providing sophisticated yet well-defined challenges for artificial intelligence practitioners. We continue this tradition by proposing the game of Hanabi as a new challenge domain with novel problems that arise from its combination of purely cooperative gameplay with two to five players and imperfect information. In particular, we argue that Hanabi elevates reasoning about the beliefs and intentions of other agents to the foreground. We believe developing novel techniques for such theory of mind reasoning will not only be crucial for success in Hanabi, but also in broader collaborative efforts, especially those with human partners. To facilitate future research, we introduce the open-source Hanabi Learning Environment, propose an experimental framework for the research community to evaluate algorithmic advances, and assess the performance of current state-of-the-art techniques.

1. Introduction

Throughout human societies, people engage in a wide range of activities with a diversity of other people. These multi-agent interactions are integral to everything from mundane daily tasks, like commuting to work, to operating the

* Corresponding author.
E-mail addresses: nbard@google.com (N. Bard), jnf@fb.com (J.N. Foerster), chandar@google.com (S. Chandar), burch@google.com (N. Burch), lanctot@google.com (M. Lanctot), songf@google.com (H.F. Song), eparisotto@cs.cmu.edu (E. Parisotto), vdumoulin@google.com (V. Dumoulin), smoitra@google.com (S. Moitra), edwardhughes@google.com (E. Hughes), iaindunning@gmail.com (I. Dunning), shibl@google.com (S. Mourad), hugolarochelle@google.com (H. Larochelle), bellemare@google.com (M.G. Bellemare), bowling@google.com (M. Bowling).
[1] Equal contribution.
[2] Work done at DeepMind.
[3] Work done at Google Brain.

https://doi.org/10.1016/j.artint.2019.103216
0004-3702/

Frontier for AI

State of the art:

Self-play	2P	3P	4P	5P
Agent type	deep RL	rule based	rule based	rule based
Average score	24.6	24.8	24.96	24.94

Ad-hoc scores are difficult to measure precisely (10 ~ 15)

Frontier for AI

Research group at NYU made diverse range of rule-based bots

Two dimensions:
risk aversion and *communicativeness*

Optimal:

- don't take much risk (0.8)
- don't talk too much (0.5)

Diverse Agents for Ad-Hoc Cooperation in Hanabi

Rodrigo Canaan
NYU
New York, USA
rodrigo.canaan@nyu.edu

Julian Togelius
NYU
New York, USA
julian.togelius@nyu.edu

Andy Nealen
USC
Los Angeles, USA
anealen@cinema.usc.edu

Stefan Menzel
HRI Europe GmbH
Offenbach, Germany
stefan.menzel@honda-ri.de

Abstract—In complex scenarios where a model of other actors is necessary to predict and interpret their actions, it is often desirable that the model works well with a wide variety of previously unknown actors. Hanabi is a card game that brings the problem of modeling other players to the forefront, but there is no agreement on how to best generate a pool of agents to use as partners in ad-hoc cooperation evaluation. This paper proposes Quality Diversity algorithms as a promising class of algorithms to generate populations for this purpose and shows an initial implementation of an agent generator based on this class. We also propose metrics that can be used to compare such generators, and how the proposed generator could be leveraged to help build adaptive agents for the game.

I. Introduction

Traditionally, research into Artificial Intelligence agents for playing games has focused on competitive, perfect information games, such as Checkers [1], Chess [2] and Go [3]. Cooperative games with imperfect information are an interesting research topic not only due to the added challenges posed to researchers, but also because many modern industrial and commercial applications can be seen as examples of cooperation between humans and machines in order to achieve a mutual goal in an uncertain environment. In previous work [4], we surveyed metrics and open problems relating to co-creative and mixed initiative systems, and argued that games, especially cooperative games, are an ideal research platform to address some of these issues.

Agent modeling is one of the main features of co-creative and mixed initiative systems identified in this survey, and cooperative games face, to a larger extent than competitive games, the problem of modelling other agents (human players or another AI agent). Usually, the main challenge in agent modeling is about *predicting* the future actions of other agents, but in our chosen domain it is also important to *interpret* observed actions and infer what they might imply about hidden features of the world. In essence, our agents should be able to represent the mental state of other actors and see the world from their perspective. This ability has also been referred to as having a theory of mind [5].

In this context, interacting with agents for which a model is known in advance is a very different problem than interacting agents for which no such model is known. When playing with known agents, a number of assumptions, or conventions, can be taken for granted, but when playing with *Ad-Hoc teammates* [6]–[8], we need either adaptive strategies that learn and leverage a model of the other players on the fly or non-adaptive strategies that play well with a wide range of partners, despite using the same policy for all of them.

One issue that emerges when dealing with *Ad-Hoc cooperative* scenarios is how to evaluate agents that play in such a setting. A typical approach is to specify a *pool* of agents that we want to be able to play well with. However, if this pool is known in advance, then *challenger* agents can be over specialized towards this specific pool, leading to behavior that might not generalize to other teammates. This typically requires the pool to be kept secret, which can lead to issues of reproducibility.

Another alternative, which we favor, is to use some stochastic method to generate a pool where agents display enough variety that different strategies are needed to play well with all of them. It is also desirable that the method produces agents with enough variety each time it is run, so it can be reused for multiple experiments without relying on secrecy and without making it too easy to over specialize agents towards it.

In this paper we discuss characteristics that would be desirable in a method for generating pools of agents that can be used to evaluate other agents in ad-hoc cooperation settings. We propose metrics that can be used to characterize these generators and implement a generator of agents using MAP-Elites [9], a Quality Diversity algorithm that optimizes towards a set of behaviorally diverse, high quality individuals, using a rule-based representation of Hanabi agents. We hope this method will help us better evaluate the challenger agents we want to develop in the future and help others to better evaluate their own agents.

II. Related work

A. Hanabi: the game

Hanabi is a cooperative card game designed by Antoine Bauza and has won the prestigious *Spiel des Jahres* award for tabletop games in 2013. It is played by groups of 2-5 players who try to play stacks of cards in correct order of rank or value (from 1 to 5) for each of the five colors in the game (R, R, Y, W and G). Players play with the contents of their hands facing outwards, so that each player sees the cards every other player has, but not their own cards. The group can only communicate through information (or hint) actions, which allow the current player to select another player and

978-1-7281-1884-0/19/$31.00 ©2019 IEEE

TU/e

Frontier for AI

Facebook AI Research developed agents called B.A.D. and S.A.D.

Keeps track of beliefs and updates them with Bayes' theorem.

Trick during training phase:

- tell the move you actually wanted to do,
- then explore

arXiv:1912.02288v1 [cs.AI] 4 Dec 2019

SIMPLIFIED ACTION DECODER FOR DEEP MULTI-AGENT REINFORCEMENT LEARNING

Hengyuan Hu, Jakob N Foerster
Facebook AI Research, CA, USA
{hengyuan,jakef}fb.com

ABSTRACT

In recent years we have seen fast progress on a number of benchmark problems in AI, with modern methods achieving near or super human performance in Go, Poker and Dota. One common aspect of all of these challenges is that they are by design *adversarial*or, technically speaking, zero-sum. In contrast to these settings, success in the real world commonly requires humans to collaborate and communicate with others, in settings that are, at least partially, cooperative. In the last year, the card game *Hanabi* has been established as a new benchmark environment for AI to fill this gap. In particular, Hanabi is interesting to humans since it is entirely focused on *theory of mind*, i.e., the ability to effectively reason over the intentions, beliefs and point of view of other agents when observing their actions. Learning to be informative when observed by others is an interesting challenge for Reinforcement Learning (RL): Fundamentally, RL requires agents to explore in order to discover good policies. However, when done naively, this randomness will inherently make their actions less informative to others during training. We present a new deep multi-agent RL method, the *Simplified Action Decoder* (SAD), which resolves this contradiction exploiting the centralized training phase. During training SAD allows other agents to not only observe the (*exploitative*) action chosen, but agents instead also observe the *greedy* action of their team mates. By combining this simple intuition with best practices for multi-agent learning, SAD establishes a new SOTA for learning methods for 2-5 players on the self-play part of the Hanabi challenge. Our ablations show the contributions of SAD compared with the best practice components. All of our code and trained agents are available at https://github.com/facebookresearch/Hanabi_SAD.

1 INTRODUCTION

Humans are highly social creatures and spend vast amounts of time coordinating, collaborating and communicating with others. In contrast to these, at least partially, cooperative settings most progress on AI in games has been in zero-sum games where agents compete against each other, typically rendering communication futile. This includes examples such as Go (Silver et al., 2016; 2017; 2018), poker (Brown & Sandholm, 2017; Moravčík et al., 2017; Brown & Sandholm, 2019) and chess (Campbell et al., 2002).

This narrow focus is unfortunate, since communication and coordination require unique abilities. In order to enable smooth and efficient social interactions of groups of people, it is commonly required to reason over the intents, points of views and beliefs of other agents from observing their actions. For example, a driver can reasonably infer that if a truck in front of them is slowing down when approaching an intersection, then there is likely an obstacle ahead. Furthermore, humans are both able to interpret the actions of others and can act in a way that is informative when their actions are being observed by others, capabilities that are commonly called *theory of Mind* (ToM, (Baker et al., 2017). Importantly, in order to carry out this kind of reasoning, an agent needs to consider why a *given* action is taken and what this decision indicates about the state of the world. Simply *observing* what other agents are doing is not sufficient.

While these abilities are particularly relevant in partially observable, fully cooperative multi-agent settings, ToM reasoning clearly matters in a variety of real world scenarios. For example, autonomous

1