Formatting Instructions for NIPS 2015

David S. Hippocampus*
Department of Computer Science
Cranberry-Lemon University
Pittsburgh, PA 15213
hippo@cs.cranberry-lemon.edu

Coauthor
Affiliation
Address
email

Coauthor
Affiliation
Address
email

Coauthor
Affiliation
Address
email

Coauthor
Affiliation
Address
email
(if needed)

Abstract

Abstract here.

1 Introduction

Stephanie

Basic hanabi description here, then talk about general applicability of cooperative AI

Using machine learning to play adversarial games is a well-studied topic. We believe that using machine learning to play a cooperative game will bring about new questions. In Hanabi, for example, the emphasis on cooperation means that its advantageous for all players to agree on certain protocols beforehand. One question that wed like to answer in this project is whether, given a human player that is already following a certain protocol, an AI can be automatically taught to follow the same protocol.

Another question is whether its possible to simultaneously train multiple models. The most obvious way to train a model may be to simulate games played with an initial, hardcoded policy, which has deterministic behavior and can therefore act as part of the environment. We believe this will work well for a two-player game, but is unlikely to scale to more players, when more coordination is needed. As part of this project, we would also like to explore the effect of simultaneously training two or more models, possibly with one hard-coded player.

2 Background

Sagar

+RL

Actually hanabi here

Hanabi [1, 2] is a cooperative card game in which two or more players act on partial information towards a common goal. Each player can see the cards of all other players, but not their own. On

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Name	Colors	1s, 2s ... 5s	Info Tokens	Fuses	Hand Size
Mini	3	[2, 2, 1, 0, 0]	6	3	3
Medium	4	[3, 2, 2, 1, 0]	8	4	4
Regular	5	[3, 2, 2, 2, 1]	8	4	5

Figure 1: Hanabi Game Sizes

each turn, a player can play a card (without looking at it), discard a card, or give information to another player. The goal is to play as many cards as possible in a specified order.

Figure 1 shows 3 sizes of hanabi that we use in our experiments.

3 Approach

Here, we describe the various components we implemented to play and learn Hanabi. Section 4 compares the performance of the different state representations, reward functions, algorithms, and training methods.

3.1 Overview

Michael

Talk about how the following pieces fit together + OpenAI

Stephanie

3.1.1 Space reps and Reward Functions

1) Nested

2) Flattened

Talk about difficulty in fixed state rep/action mapping

3.2 Training

Make algs/heuristics subsections here, talk about different combos

3.2.1 Algorithms

3.2.2 TRPO

[3]

3.2.3 VPG (if space)

3.2.4 CEM (if space)

3.2.5 CMA-ES (if space)

3.3 Heuristics

We implemented hardcoded heuristic-based Hanabi players to see if we could guide the previously discussed algorithms into learning a particular strategy. Below we discuss two of these heuristic-based players.

3.3.1 Heuristic

Stephanie

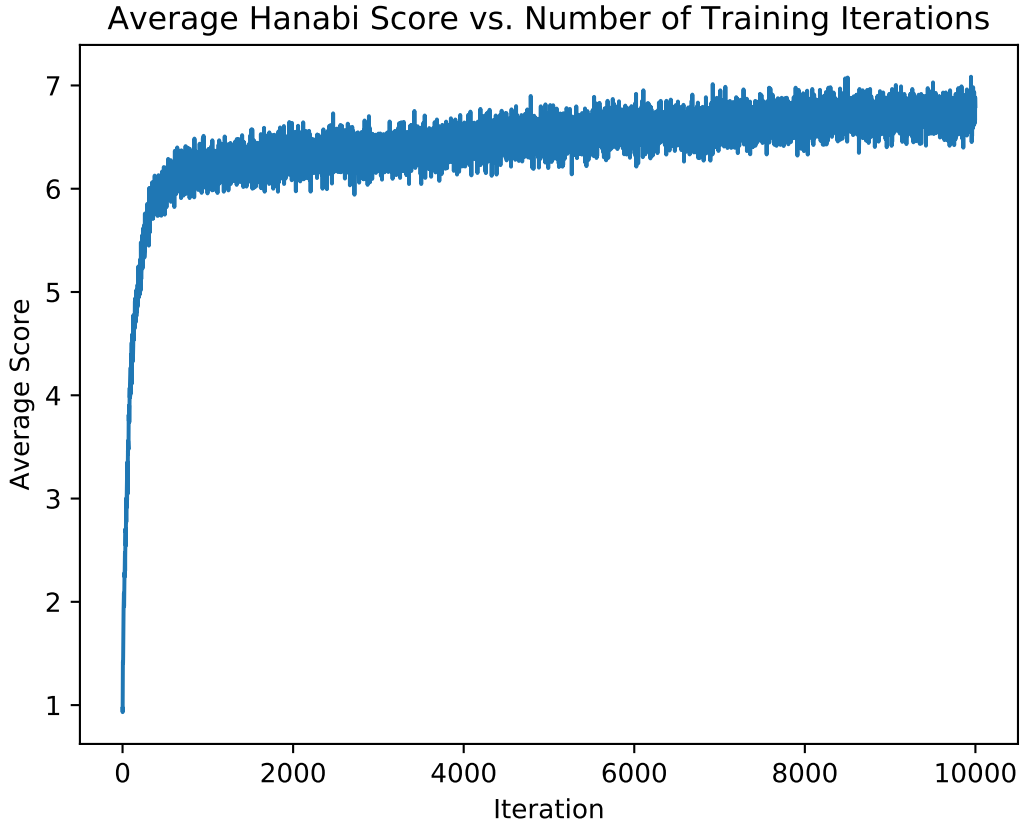

Figure 2:

3.3.2 SimpleHeuristic

The SimpleHeuristic player attempts to mimic the strategy that our best TRPO-trained-on-itself player learns for mini-hanabi. Each turn, it tries the following, prioritizing lower numbered cards first:

- Give info about cards of the given number if there are any that the other player does not know about
- If a card of the given number is already played for each color, discard any card of that number that we know about in our hand
- Play a card of the given number.

If none of these conditions are satisfied, for example if the other player knows the number for each of their cards and the heuristic player knows nothing about their own cards, the player tries to play the zeroeth card in its hand. We avoid discarding as much as possible, since every discard after the two costs one point in two player mini-hanabi.

4 Evaluation

Plots:

Mini TRPO small iters, hyperparam tuning NN size 8x8 -¿ 256x256 by 2s + more layers step size -¿ diff fixed + dynamic batch size -¿ fix batch size * iter, scale them

Mention other algs -establish that TRPO is the best

TRPO vs heuristics

TRPO many iters mini, medium, regular

TRPO mini histogram of scores

vs other bots

TRPO strategy

Human vs Human Human vs TRPO

5 Conclusion

References

References follow the acknowledgments. Use unnumbered third level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to 'small' (9-point) when listing the references. **Remember that this year you can use a ninth page as long as it contains *only* cited references.**

References

[1] Hanabi — board game — boardgamegeek. `https://boardgamegeek.com/boardgame/98778/hanabi`. Accessed: 2017-4-30.

[2] Hanabi (card game). `https://en.wikipedia.org/wiki/Hanabi_(card_game)`. Accessed: 2017-4-30.

[3] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel. Trust region policy optimization. *CoRR*, abs/1502.05477, 2015.