RL Generalization in a Theory of Mind Game Through a Sleep Metaphor (Student Abstract)

Tyler Malloy,[1 2] **Tim Klinger,**[2] **Miao Liu,**[2] **Gerald Tesauro,**[2] **Matthew Riemer**[2] **and Chris R. Sims**[1]

[1] Rensselaer Polytechnic Institute Department of Cognitive Science, Troy, NY, USA
[2] IBM T.J Watson AI Research, Yorktown Heights, NY, USA
{mallot, simsc3}@rpi.edu {tklinger, mdriemer, gtesauro}@us.ibm.com, miao.liu1@ibm.com

Abstract

Training agents to learn efficiently in multi-agent environments can benefit from the explicit modelling of other agent's beliefs, especially in complex limited-information games such as the Hanabi card game. However, generalization is also highly relevant to performance in these games, though model comparisons at large training timescales can be difficult. In this work, we address this by introducing a novel model trained using a sleep metaphor on a reduced complexity version of the Hanabi game. This sleep metaphor consists an altered training regiment, as well as an information-theoretic constraint on the agent's policy. Results from experimentation demonstrate improved performance through this sleep-metaphor method, and provide a promising motivation for using similar techniques in more complex methods that incorporate explicit models of other agent's beliefs.

Introduction

Hanabi has been suggested as a potential area of research in artificial intelligence due to the unique properties of the learning environment which require human players to predict the beliefs of other agent's through understanding the actions they perform (Bard et al. 2020). This makes the challenge of the game considerably easier for humans, who have a natural ability to understand the actions of others. One method of improving performance of RL agents in the Hanabi environment would be to try and explicitly model this human faculty and incorporate it into the model. However, the method described in this paper seeks to improve performance over existing methods in RL by training agents to have more generalizeable behaviour, instead of explicitly representing the beliefs of other agents. Results from these experiments show that this is possible through augmenting training with periods of either a sleep or dreaming metaphor.

The goal of the Hanabi card game is to play each card in order from 1 to 5 onto the board which consists of piles sorted by color. This is done by either playing a card onto the board, discarding a card, or giving a player information, for a more complete explanation of the rules see the supplementary material. Recent work in the Hanabi card game has focused on the traditional rules of the game (Bard et al. 2020), but in order to compare many different models we consider a reduced version with 2 players, card colors, values, and cards in hand, instead of 5 for each value. This is done to allow for a better investigation of learning curves of agents, as well as the stability of performance once achieving an optimal strategy, as in previous works attempts at training RL agents have not lead to optimal performance in the full task.

Capacity-Limited RL

The method described in this paper is based on altering the traditional reward maximization method in RL by introducing a penalty to the reward observed by the agent based on the informational complexity of their policy as follows:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim p_\pi}\big[r(s_t,a_t)\big] - \beta\mathcal{I}(\pi(a_t|s_t)) \quad \text{(1a)}$$

$$\mathcal{I}(\pi(a_t|s_t)) = \mathbb{E}_{s_t,a_t}\left[-\log\left[\frac{\pi(a_t|s_t)}{\pi_a(a_t)}\right]\right] \quad \text{(1b)}$$

With $\mathcal{I}(\pi(a_t|s_t)$ representing the informational complexity of the agent's policy. Where Eq. 1b describes the mutual information between the state of the agent and the action to perform, as defined by the agent's policy.

This learning objective in Eq (1a) defines the Capacity-Limited approach, a general method of reinforcement learning that can be applied onto any existing RL method. In this paper, it is applied onto the Advantage Actor Critic (A2C) (Mnih et al. 2016) to develop the Capcity-Limited Advantage Actor Critic (CLA2C) model. The motivation for this alteration is to encourage the reuse of learned behaviour across different states of the environment, which has been shown to improve the generalizability of learned behaviour in RL agents (Lerch and Sims 2019).

In Eq. (1a) the parameter β serves to balance reward and informational simplicity, with the traditional reward maximization method being preserved as β approaches zero. Results from training the CLA2C include a constant information constraint with $\beta = 0.025$.

Sleep Metaphor

Recently it has been suggested that one function of dreaming in humans is in improving the generalizability of learned behaviour (Hoel 2020). This interpretation views the randomness and non-sequetorial nature of human dreaming to be an

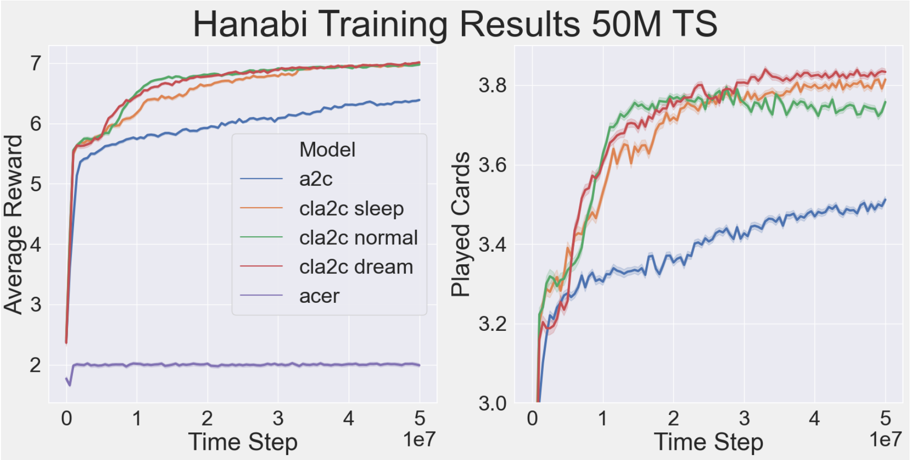

Figure 1: Average Reward and Number of Played Cards for A2C (blue), CLA2C (green), CLA2C with sleep training (orange), CLA2C with dream training (red), and ACER in purple. All results are in the Hanabi game with 2 colors, numbers, and players

effort by the brain to avoid over-fitting learned behaviour on the sparse experience of waking life.

The way that the sleep metaphor is instantiated in the model described in this paper is by augmenting training with periods of either ‘sleep’ or ‘dreaming’. In the sleep condition the capacity limited learning objective in Eq (1a) is only applied for short periods in between training. The dream state consists of the same information capacity as in the previous method, with the difference being that games starting positions are randomly distributed board, as if starting the game in the middle of an already existing game.

Experiment Results

Because the CLA2C method is built off the existing A2C method, it serves as the main point of comparison for performance. The sleep/dream method shares a close theoretical justification with experience replay methods, and thus the ACER model (Wang et al. 2016) was also trained in the Hanabi environment.

Firstly, we can see that the CLA2C model with sleeping and dreaming have similar performances, with dreaming having a slightly faster improvement on reward, but the same average number of cards played. All capacity-limited methods have improved performance on reward compared to the A2C and ACER models.

Additionally, there is a slight decrease in the number of played cards by the normal CLA2C model. This can be explained by considering the motivation for using a capacity-limit which could be described as learning a useful marginal policy to use as a prior, penalizing actions that deviate too far from that prior. Because the agent is rewarded for giving other players information, the CLA2C agent likely quickly learns the value of giving information to other players in the Hanabi game. However, too high a preference for giving information may slightly decrease the probability of playing each card. This suggests that using the sleep or dream training methodology improves generalization and stability by avoiding this decrease in number of cards played.

Discussion

There has been much interest in methods of reinforcement learning that incorporate more complex models that explicitly represent certain aspects of a learning agent that would be desirable for different environments. However, when analyzing the usefulness of a new approach in RL such as the capacity-limited and slep-metaphor method, it is often insightful to apply it standard methods. The results detailed in this paper demonstrate that the capacity-limited approach and the sleep/dream metaphor can be used to improve the generalizability of learned behaviour in this complex learning task without representing the behaviour of other players. Hopefully, further investigation of these methods alongside more complex techniques for representing other player’s behaviour will show similar improvements in performance.

References

Bard, N.; Foerster, J. N.; Chandar, S.; Burch, N.; Lanctot, M.; Song, H. F.; Parisotto, E.; Dumoulin, V.; Moitra, S.; Hughes, E.; et al. 2020. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence* 280: 103216.

Hoel, E. 2020. The Overfitted Brain: Dreams evolved to assist generalization. *arXiv preprint arXiv:2007.09560* .

Lerch, R.; and Sims, C. 2019. Rate-distortion theory and computationally rational reinforcement learning. *Proceedings of Reinforcement Learning and Decision Making (RLDM) 2019* 7–10.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937.

Wang, Z.; Bapst, V.; Heess, N.; Mnih, V.; Munos, R.; Kavukcuoglu, K.; and de Freitas, N. 2016. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224* .