Q-Learning in Shum-Style Stag Hunt

Adrian Manhey
amanhey.edu
Occidental College

1 Introduction and Problem Context

In the machine learning field, reinforcement learning (RL) has received wide recognition for its ability to solve complex problems, such as a set of Atari games where it achieved superhuman performance in certain games [6]. One of the areas of interest in reinforcement learning is in the application in multi-agent environments, such as stag hunt [12], a stochastic prisoner's dilemma-style game where the goal is to accumulate the highest number of points by capturing high-value mobile targets (stags) or low-value static targets (hares). The stochastic iterated prisoner's dilemma is a kind of iterated prisoner's dilemma game where the strategies of the players are specified in terms of cooperation probabilities [5]. In this task, agents control hunters which can hunt stags or hares, but stags require two or more agents to capture them while hares only require one.

Figure 1. One of the nine stag-hunt scenarios. Circles represent hunters' position at a given time step. Dotted lines represent motion. Figure adapted from Rabkina et al. (2019).

Figure 1 is an example stag-hunt scenario [9] where a pair of agents (A and B) cooperate to capture a stag and another agent captures a hare individually. The circles represent a hunter's position at a given time-stamp and the dotted lines represent movement. A key element of this game is the collaboration between agents which can be modeled using various methods. One such method is a type of machine learning called Q-learning, which was used as the winning model in the Malmo Collaborative AI Challenge, hosted by Microsoft in 2017, a version of stag hunt where the agent was given two choices: 1) catch the pig by trapping it in a corner and receiving 25 points or 2) giving up and receive 5 points [3]. However, this solutions has limitations in how it could be extended beyond the Pig Chase task since the Q-learning model would need to be retrained for each new task, which is nontrivial, and that the secondary agent would choose either random actions or followed a heuristic search. A rendering of the Malmo-Challenge environment can be seen in Figure 2.

Two of the solutions to these issues involved using a Bayesian model [11] and Analogical Theory of Mind (AToM) [9], each of which will be discussed more in Section 2. The AToM solution, which is the parent research project to this project, posited that AToM reasoning would allow agents to infer cooperation intentions of other agents and make predictions about other agents' future actions, leading to a model with a higher level of generalizability. However, to test this data would need to be collected on the performance of that model in comparison to different implementations, such as a RL model. This project aims to create a RL mode for comparison for the AToM model using Q-learning with four additional factors to Pig Chase: 1) replacing the low-reward target of quitting with hares, 2) implementing the game with three hunters, two hares, and two stags, 3) the ability for the secondary agent to seek the low-reward targets and 4) using a 5x7 grid world.

These changes follow the spatial stag-hunt domain from Shum et al.'s model and allow a direct comparison between the two models, as well as data to be collected on human inference of the cooperation of agents. In addition to the implementation of the Q-learning model, a reach-goal for the project is to gather data in a small study where participants would view episodes of the model playing stag hunt and inferring whether they perceive the agents cooperating together or not.

2 Technical Background

The implementation of the Q-learning model will be done with OpenAI's Gym environment, designed for "developing and comparing reinforcement learning algorithms." [8, 2] This environment provides a set of pre-made environments to reference as well as the ability to create a custom environment. Since this Shum-style implementation is a grid-based environment with the only actions being movement in the four cardinal directions, it will have a discrete observable- and action-spaces. In the Shum et al. domain there are 9 maps that are used, so the observable-space will change depending on which map is chosen. That set must be taken into account so that either the project uses a single map or a set of Q-learning models for different maps. The action-space, however, will be consistent throughout each map. Additionally, there were only time-steps in the implementation, which will remain true in the implementation of this

project.

In terms of the model, a table of state and action pairs (s, a) with an associated Q-value, representing their "quality," will be defined for each maps in the set. Q-values are initialized to an arbitrary value and as the agent explores the environment they are updated. The method for updating a Q-value consists of two parameters: $\alpha \in (0, 1]$ is the learning rate and $\gamma \in [0, 1]$ is the discount factor, which makes short-term or long-term rewards more valuable [4]. The equation for this relationship can be written as

$$Q(s, a) = (1 - \alpha)Q(s, a) + \alpha(r + \gamma \text{max}Q_a(s', A)),$$

where r is the reward of an action, s' is the next state, and A is all the possible actions that can be taken []. Additionally the algorithm uses ϵ, which can be thought of the exploration parameter. This tells the algorithm how often to pick a random action instead of the current most beneficial one from the Q-table. In sum, we are learning the proper action by taking the old Q-value and adding the learned value of the immediate reward and possible future rewards. This algorithm also has the potential to be optimized by modifying the values of the parameters either before or during training. For example, the learning rate can be decreased as the knowledge base increases and the exploration parameter can be decreased as the number of trials increase.

The secondary agents will be implemented using a heuristic search algorithm, such as A-star. The goal of these secondary agents are to mimic basic human-play, to some reasonable degree, to train the model to be able to cooperate with a human without require large amounts of human-play data, which would be infeasible to collect for this project. The heuristic for these agents will operate around distance (e.g. distance of other hunters to targets, a hunters distance to a target compared to other hunters).

3 Prior Work

The two solutions touched on earlier aim to solve the problems of cooperation inference and action prediction in the game. From the perspective of one agent, there are two questions to be answered: 1) are other agents trying to cooperate with me to catch a stag and 2) what behavior will they exhibit if they are/are not? Shum et al.'s model is based on Bayesian inference with an explicit causal model, which map different behaviors to different states of the group of hunters. Due to the explicit team hierarchy this model does not require training and was found to have strong correlations with human predictions in the experiments conducted to address the previous two questions. Their contribution was a novel representation for extending single-agent generative models of action understanding to the to multi-agent interaction. However, one of the limitations of this implementation is that it only works well for small groups, since as the number of agents grows the explicit team hierarchy also grows.

Alternatively, Rabkina et al. proposed a solution called Analogical Theory of Mind. This process involves social reasoning called Theory of Mind [10], where an entity makes inferences about another's mental state (knowledge, beliefs, preferences, desires, goals, and intentions). The AToM model posits that a combination of analogical processes and feedback lead to the development of theory of mind reasoning without an underlying model of cooperation. The AToM model is more accurate than the Bayesian model at most time stamps and does not require a predefined underlying hierarchy of team cooperation. Furthermore, the ability to make predictions about the future actions of agents is an additional effect of reasoning about an agent's cooperation with others.

Figure 2. Microsoft Malmo Platform PigChase environment. Left: actual 3D rendering in the Mamlo Platform. Right: symbolic representation of agent and pig on the map.

In addition to these models that utilize cognitive architectures for their reasoning, there are some models that use deep learning instead as part of the growing application of deep learning models in RL [7]. Arulkumaran et al. created a model that used the REINFORCE algorithm to create a 4 layer convolutional neural network [1, 13]. They began by "pre-training" on an PigChase Replica Environment, a domain which approximated the Malmo-Challenge environment and was used to generate large batches of episodes for training. Then they trained in the Malmo-Challenge environment using batches of 128 episodes and used the mean discounted return, for that particular time step in the game, as their evaluation metric. Their resulting model was able to outperform a A-start heuristic, as expected, and solve the puzzle. The significant question to consider from such a research project is the transferability of their model to different domains and how it's flexibility and performance compares to different models, such as Q-learning.

4 Methods and Evaluation Metrics

This project aims to create a RL mode for comparison for the AToM model using Q-learning with four additional factors to Pig Chase: 1) replacing the low-reward target of quitting with hares, 2) implementing the game with three hunters, two hares, and two stags, 3) the ability for the secondary agent to seek the low-reward targets and 4) using a 5x7 grid world.

First idea would be to implement a Q-learning algorithm like the taxi tutorial.

5 Ethical Considerations

Think about what the research is and the goals of the project, particularly the social goals. If the goals are still dealing with people, there will be various cultural and racial considerations to be made. This is built on the assumption that the project is focused on robot-human interaction. However, if this is built on robot-robot cooperation then it doesn't matter. You can also make a simplified model of a person, defined as a sequence of events that can be skipped through. Also need to consider how that agent is going to perform with the desired player given the testing player, if it trains with a RL agent can it play with a human agent.

6 Timeline

The first step is going to be doing more research around reinforcement learning, especially in stag-hunt. So far there have been quite a few resources for both single and mult-agent stag-hunt using reinforcement learning so finding which algorithm will work best for this project will be key. After the Occidental College URC program begins, I plan to begin doing this research in tandem to my other role on the parent project. Most of the summer will mainly consist of this research so that in the fall the project has a clear direction. In May I plan to learn more about the existing models of the project and how the new implementation may account for some gap in functionality or theory. June will likely then be reserved for researching models I would want to build myself, which involves loosely designing the model I choose in terms of high-level implementation. This phase of the project will be building the prior work that has been done in terms of stag-hunt, reinforcement learning, or multi-agent games. In July I plan to continue designing the model and developing the kind of technical information I may need. Additionally, this time period of high- to mid-level design would provide a good opportunity to start putting together a list of resources needed by the project, i.e. are there any computational tools I might need to request in the fall? The semester starts in August so I would like to finish the design of the model in order to have it ready to implement in September. This phase revolves around creating the specific methods I'm using in my implementation. In September I plan to finish implementing the model and start trying to improve it, e.g. optimizing the parameters. October will consist of evaluating the model and gaining insight into the performance of the model and how well it meets the goals of my comprehensive project and the parent project. The last two months will involve any needed refactoring and preparing for presentations.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque augue neque, tempor sed tincidunt at, condimentum sed tortor. Aliquam cursus consectetur ex eu dictum. Curabitur rhoncus dapibus sem, et dapibus arcu semper a. Sed massa eros, mollis aliquet lacus a, pretium molestie sapien. Suspendisse egestas varius ultricies. Integer cursus elit lorem, ut tempus diam convallis vel. Mauris convallis mauris est. Sed at ex venenatis, viverra enim eget, porta nisi.

Phasellus at condimentum tellus, ac cursus dolor. Donec hendrerit mattis ultrices. Vestibulum et mi sed nibh suscipit tincidunt sit amet at ipsum. Pellentesque dolor lorem, feugiat vel cursus et, efficitur non orci. Nulla lacus est, gravida ut viverra in, malesuada sed odio. Nam volutpat neque et sollicitudin imperdiet. Integer auctor viverra vulputate. Aenean ut libero et sem imperdiet luctus et id orci. Mauris nec arcu in diam faucibus euismod tempus in quam. In hac habitasse platea dictumst. Curabitur quis sem ac magna iaculis dictum. Morbi vel suscipit sem. Sed vitae aliquet ex.

Nulla tempor pellentesque massa tristique pellentesque. In nec euismod nibh, a semper arcu. Donec nec cursus nibh, et semper nibh. Nunc sagittis nisl ante, eu porta orci blandit egestas. Ut consequat leo at massa vestibulum, at cursus libero tristique. Vivamus semper tincidunt tellus, et aliquet magna scelerisque eget. Morbi pharetra feugiat est, lacinia lobortis sapien sagittis in. Aliquam erat volutpat. Vestibulum fermentum fringilla eros, et ullamcorper felis hendrerit feugiat. Suspendisse sed lobortis elit. Donec quis pharetra lorem.

Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Etiam aliquet lorem at augue faucibus, eu iaculis leo accumsan. Donec ullamcorper elit nisi, at tincidunt turpis suscipit pharetra. Donec auctor erat sem, ut ultricies lectus egestas sed. Nunc mattis congue sapien, id varius augue. Donec non risus vitae dolor consequat laoreet ut in mi. Fusce molestie tellus a tristique volutpat. Aliquam felis justo, iaculis sed metus et, consequat imperdiet ipsum. Integer tristique ac ipsum quis eleifend. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nulla facilisi. In sem nulla, fermentum at elit eget, bibendum tempor felis.

Maecenas ullamcorper nibh sed purus ullamcorper, suscipit convallis velit sagittis. Duis neque est, imperdiet in gravida et, accumsan a eros. Sed placerat tristique risus ac maximus. Cras at lorem eu purus rhoncus suscipit. Etiam rhoncus laoreet vehicula. Vivamus aliquet tristique diam, id pellentesque eros ornare a. Nullam fermentum nec tellus venenatis ultricies. Praesent laoreet nec urna quis accumsan. Cras pulvinar sapien ac justo malesuada, nec scelerisque nunc vulputate. Mauris non elit id ante convallis scelerisque nec vel odio. Maecenas mollis elementum eros in gravida. Mauris luctus pellentesque lacus, sit amet mattis quam ultricies congue. Etiam et odio justo.

Maecenas laoreet, ex at volutpat ultricies, odio nisi fermentum tellus, quis porta odio felis eget orci. Donec interdum fringilla elementum. Phasellus aliquam leo et leo

rutrum commodo. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Maecenas mauris arcu, blandit tincidunt bibendum venenatis, faucibus ut mauris. Suspendisse vulputate nunc sit amet risus placerat, interdum pretium magna hendrerit. Praesent vel bibendum sem. Praesent imperdiet id ex eget gravida. In volutpat dui eget massa consequat, nec ultricies felis tempor. Duis feugiat mauris sed tellus vestibulum feugiat. Pellentesque ante orci, elementum sit amet leo posuere, luctus congue justo. Pellentesque vulputate cursus ligula, in blandit risus tempus id.

Duis a neque aliquet, pharetra risus vitae, sodales turpis. Nullam non metus fermentum, venenatis quam molestie, fringilla justo. Sed at malesuada nisi. Curabitur aliquet neque ante, vitae fermentum metus ultricies in. Sed pharetra condimentum ex vel pharetra. Cras tempor, purus non placerat dapibus, velit mi ultrices massa, et elementum dolor tortor et dolor. Donec sed arcu ac lectus iaculis congue quis eget mi. Fusce eleifend aliquet ligula, id porttitor felis consectetur a. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque sit amet ullamcorper tortor. Vivamus feugiat nisl id metus interdum, vel tincidunt lacus volutpat.

Aliquam varius eu lectus tempor ornare. In semper auctor urna, ut auctor odio faucibus non. Nunc leo eros, hendrerit id elementum vitae, imperdiet et nisi. In a pretium lacus, in tincidunt orci. Aenean et dui urna. Donec vulputate nibh id tortor bibendum viverra. In iaculis non nibh quis pulvinar. Integer id sagittis diam.

Sed augue quam, fermentum a quam ut, tincidunt pretium nunc. Aenean auctor, tortor ac porta convallis, ligula nisl consequat nulla, sit amet fermentum diam nulla quis metus. Nunc tellus erat, blandit non ante ultrices, tempus molestie sapien. Suspendisse potenti. Donec hendrerit viverra nibh eu consequat. Vivamus lectus lacus, imperdiet vitae lorem non, pellentesque bibendum nunc. Nam ultrices eros ut felis consequat viverra. Donec congue vestibulum molestie. Curabitur id porttitor leo. Cras erat metus, luctus eget tellus eu, dignissim sodales enim.

Duis fringilla dapibus arcu, ut tristique enim hendrerit at. Aliquam erat volutpat. Phasellus quis tortor et nulla dapibus tristique et sit amet metus. Mauris nec faucibus leo, et imperdiet purus. Mauris sit amet sapien ut enim pellentesque blandit. Nunc luctus libero magna, ut ullamcorper nibh sodales vel. Cras sodales scelerisque erat, vel condimentum est tristique at. Nam quis dolor turpis. Fusce varius velit in magna molestie bibendum. Nulla facilisi. Nullam maximus fermentum dapibus. Vivamus eu rutrum risus, a dignissim nunc. Nunc non purus dictum, commodo mauris sit amet, sollicitudin tortor. Duis ornare nulla eu mi vulputate, at feugiat nunc condimentum.

Duis malesuada lacus ac mi viverra, a venenatis elit faucibus. Pellentesque congue ipsum non scelerisque lobortis. Etiam eu dapibus velit, at tempus tortor. Suspendisse in elit ac ligula consequat hendrerit vel sed sem. Duis a imperdiet erat. Nulla facilisi. Etiam ultricies urna vitae mauris posuere, eget suscipit justo rutrum. Pellentesque molestie, mauris sed commodo sollicitudin, est leo imperdiet justo, a tempor dui nisl a tellus. Nam erat ligula, efficitur non mattis ac, pharetra a justo. Proin semper arcu ut nibh interdum aliquet. Aenean lorem neque, aliquam in suscipit quis, dignissim pellentesque arcu. Quisque tincidunt, sem at imperdiet commodo, lorem nisl ullamcorper nisl, sit amet porttitor libero purus vitae nibh. Vestibulum ut augue in justo condimentum iaculis. Fusce et ligula velit. Mauris condimentum cursus mollis.

Nulla vel ipsum mollis tortor cursus hendrerit. Donec rhoncus et nisl ut accumsan. Donec eget sem sagittis, convallis neque vel, posuere nisi. Pellentesque mollis, metus eu ullamcorper porttitor, magna purus vestibulum metus, sed semper ex velit in libero. Proin augue mi, hendrerit sed pellentesque eget, ullamcorper vitae ante. Nulla et ipsum non nisi mollis efficitur vel in lectus. Maecenas non felis ut lorem finibus tristique. Maecenas blandit mi magna, mollis dictum lorem pretium sit amet.

Nullam lorem est, imperdiet id ullamcorper eget, dapibus vitae ante. Vivamus ullamcorper neque sed lacus tincidunt molestie. Aliquam sodales felis odio, a pretium odio maximus sed. Vestibulum luctus enim a nisi facilisis, eget tincidunt tellus convallis. Phasellus ut tortor tellus. Cras placerat tempor elit, vitae luctus lectus venenatis mollis. Quisque porta semper ex quis faucibus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Maecenas tempus mattis eros, sed ultricies metus. Vivamus pulvinar, tellus id elementum accumsan, tortor nisi ullamcorper augue, eget scelerisque arcu turpis imperdiet eros. Pellentesque et semper dolor. Cras vitae viverra dolor, vitae tristique urna. Sed iaculis eros et elit sodales, a gravida elit dignissim. Curabitur sem lacus, malesuada vitae eros sed, egestas blandit risus. Morbi ultricies, urna et porta vestibulum, orci purus maximus felis, et pulvinar neque erat sit amet felis.

Nam vitae fringilla metus, vitae pellentesque augue. Nam at rutrum nisi. Aenean gravida laoreet pharetra. Fusce eu leo sed turpis vulputate ultrices in vel ex. Sed vel mi mi. Ut ultrices, nulla vel laoreet vulputate, dolor enim efficitur velit, quis dignissim purus ante nec mi. Morbi accumsan iaculis suscipit. Proin cursus nisl non nisl eleifend, ac mollis massa viverra. Etiam fermentum elit nec ligula imperdiet, vitae efficitur diam laoreet. Maecenas in eleifend orci.

Nullam dapibus, nisl pretium venenatis dictum, diam lorem volutpat ex, maximus porttitor tellus enim quis odio. Vestibulum in tellus blandit, suscipit velit ut, ornare velit. Aliquam congue massa sit amet viverra facilisis. Morbi quis tincidunt orci. Nam fermentum tellus non felis bibendum, et convallis nunc ultrices. Class aptent taciti sociosqu ad

litora torquent per conubia nostra, per inceptos himenaeos. Suspendisse sagittis, nulla non iaculis vestibulum, nisi est efficitur lorem, at malesuada ex risus a tortor. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Vestibulum arcu tellus, eleifend semper elit eget, sagittis porttitor est. Nulla gravida rhoncus tellus at bibendum. Ut consectetur felis aliquam imperdiet posuere. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer id placerat ligula. Fusce ultricies sed tellus a posuere. Etiam nunc ante, porta sed magna sed, mollis mollis sapien. Aliquam quis viverra diam, rutrum porta tellus.

Donec eget consectetur risus. Duis mattis auctor urna, sit amet pharetra augue ullamcorper nec. Nam vitae iaculis libero. Fusce sed purus erat. Morbi vitae tincidunt magna. In volutpat tortor ac felis pretium, vitae venenatis metus consequat. Aenean dapibus, dolor in dictum varius, elit ipsum scelerisque arcu, a placerat tortor nibh a nunc. Cras tristique feugiat enim sed vehicula. Mauris dapibus condimentum metus vitae ultrices. Morbi nulla ante, mattis non nunc eget.

References

[1] Arulkumaran, Kai et al. "A Brief Survey of Deep Reinforcement Learning". In: *ArXiv* abs/1708.05866 (2017).

[2] Brockman, Greg et al. "OpenAI gym". In: *arXiv preprint arXiv:1606.01540* (2016).

[3] Hofmann, Katja. *The Malmo Collaborative AI Challenge*. Apr. 2017. URL: https://github.com/Microsoft/malmo-challenge.

[4] Kaelbling, Leslie Pack, Littman, Michael L., and Moore, Andrew W. "Reinforcement Learning: A Survey". In: *J. Artif. Intell. Res.* 4 (1996), pp. 237–285.

[5] Li, Siweli. *Strategies in the Stochastic Iterated Prisoner's Dilemma*. 2014. URL: http://math.uchicago.edu/~may/REU2014/REUPapers/Li,Siwei.pdf.

[6] Mnih, Volodymyr et al. "Playing Atari with Deep Reinforcement Learning". In: *CoRR* abs/1312.5602 (2013). arXiv: 1312.5602. URL: http://arxiv.org/abs/1312.5602.

[7] Nica, Andrei Cristian et al. "Learning to Maximize Return in a Stag Hunt Collaborative Scenario through Deep Reinforcement Learning". In: *2017 19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)* (2017), pp. 188–195.

[8] *OpenAi Gym Docs*. URL: https://gym.openai.com/.

[9] Rabkina, I. and Forbus, K. D. *Analogical Reasoning for Intent Recognition and Action Prediction in Multi-Agent Systems*. 2019.

[10] Ruhl, Charlotte. *Theory of Mind*. 2020. URL: https://www.simplypsychology.org/theory-of-mind.html.

[11] Shum, M. et al. *Theory of minds: Understanding behavior in groups through inverse planning*. July 2019.

[12] Skyrms, Brian. *The Stag Hunt*. Mar. 2001. URL: https://www.socsci.uci.edu/~bskyrms/bio/papers/StagHunt.pdf.

[13] Williams, Ronald J. "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine Learning* 8.3 (1992), pp. 229–256. DOI: 10.1007/BF00992696. URL: https://doi.org/10.1007/BF00992696.