First Year Report

Daniel Hernandez

1 Introduction

Classical approaches to training a good performing policy on a environment with multiple agents suffer from the problem of overfitting. Where learning agents become good at operating with or against themselves, but considerably drop performance when matched against other agents that act differently to those they have previously encountered. This issue is specially disastrous when one knows not of any existing good opponent strategies to test against. In the absence of a dataset of already existing agent strategies, multi agent reinforcement learning offers the posibility of safely learning a good policy in these kind of environments by trial an error. Making it possible to learn how to solve these problems by training from the experiences encountered inside simulated environments.

The focus of my PhD research is to adress an open question in Multi Agent Reinforcement Learning that helps to mitigate the overfitting problem. How can we, in a qualitatively and quantitative manner, analyze the way that different agent strategies in an environemnt influence the eventual strategy of a learning agent. The goal is that by presenting an agent with a sufficiently varied set of strategies the strategy learnt by this agent will be robust to changes in the opponent strategies.

2 Context

My Literature review is split in two documents. The first one can be found as part of this submission on separate document. The second part of my literature review can be found in my paper submission for the ICRA 2019 conference, paper which I co-authored. This paper submission can be found on a separate document.

3 Conclusion

3.1 Areas requiring further studies

After having heavily researched single agent scenarios under the framework of reinforcement learning, I am broadening my research to multi agent environments. Multi agent environments have been heavily studied in reinforcement learning, but there are other fields which also focus on multiagent scenarios and the relationship among multiple agents behaviours. These are game theory and evolutionary computation.

I will attend Daniel Kudenko's *Multi-agent Interactions & games* course to receive a formal education on game theory in order to be able to carry game theoretical research with a strong theoretical basis. My short term aim is to study the concept of mixed strategies in imperfect imformation games together with methods for learning these mixed strategies. This course features not only lectures on these topics, but also a set of lectures linking together game theory with reinforcement learning, which will allow me to transfer some of my RL knowledge into game theory.

There are many concepts that I want to cover in my upcoming RL research that have already been explored in the field of evolutionary algorithms. I will initially study the notion of a population archive, also known as an elite or a hall of fame, which is closely related to some of the ideas that I want to explore in my self-play research. It has been pointed out by many researchers that there is a lack of papers bringing together notions of evolutionary computation and reinforcement learning. There are many professors and lecturers within IGGI researching evolutionary algorithms, including my new supervisor, from whom I will be able to learn and discuss these topics for the benefit of my own research.

3.2 Relevant research methods, techniques and theoretical approaches

The main research method that will be used in my research is AI training via simulation, which is the only research method used in reinforcement learning. My experiments will track and analyze various metrics obtained from these simulation. Such metrics will be used to quantitatively determine the effects of different self-play variations on the learning of a good policy for a multi-agent scenario. This analysis will be run over a set of environments, using both traditional RL algorithms and deep RL algorithms. The environments will be zero-sum two-player markov games.

I will use Unity game engine to build my environments and its Unity ML-agents framework to extract simulation metrics. Python will be used to build the traditional RL models, and the Tensorflow library for deep RL models. The two deep RL algorithms are Deep Deterministic Policy Gradient (DDPG) algorithm and Proximal Policy Optimization (PPO).

3.3 Preliminary progres

I have already implemented a system which supports the theoretical requirements of the self-play system presented at the IGGI 2018 conference. It has been implemented as an extension of Unity's ML-agents framework. This was done using Python and Tensorflow. It lacks in formal automated testing. But its implementation has been manually tested using various test environments. I am at a stage where I can design an environment for which training an agent using self-play leads to learning robust policy, or set of policies (depending on the self-play system used).

For the traditional RL algorithms, I have coded Q-learning, First visit Monte carlo and Last Visit Monte Carlo. Other Possible candidates are Q(σ) and TD(λ). Regarding the deep RL algorithms, I already have a tested implementation of PPO. DDPG remains to be implemented.

4 Plan

My plan for my second year is devided into 3 parts: self-play research, social robotics paper and mandatory modules.

4.1 Self-play research

As introduced in the Conclusion Section, and as presented throughout the IGGI 2018 conference, I have been spending my research efforts studying the notion of self-play in reinforcement learning. This is my 2nd year plan with regards to research on self-play:

1. **Propose self-play framework**
   After reading a lot of research on algorithms that use self-play to train agents in a reinforcement learning framework, I have began devising a generalized self-play framework. This idea was presented as a poster and a workshop at the IGGI 2018 conference. I want to put more focus on developing this framework theoretically.

2. **Replicate other self-play systems from the literature under my proposed framework**
   My first set of experiments focuses on using my proposed framework to replicate self-play mechanisms which have already been independently studied in the literature. This will become the first benchmark of self-play mechanisms in the literature, and a test for the generalization potential of my framework. I have already identified which are the systems that I intend on replicating.

3. **Look for an algorithmic gap in the literature that can be filled with my framework**
   The intention behind creating a benchmark of existing algorithms is two-fold. On a technical view, by implementing various existing self-play mechanisms, I will gain technical knowledge on how to successfully build these systems on existing machine learning frameworks. On a theoretical view, by understanding these self-play systems, I will be in a good position to discover algorithmic gaps in this part of my field. Once I discover either a new algorithm, or an algorithmic improvement, I will focus on developing it theoretically. This new algorithm would be easily benchmarked against the already implemented algorithms.

4. **IJCAI 2019**
   The prestigious IJCAI 2019 conference will be held in August in Macao, China. Its call for papers ends on the 25th of February. I intend to publish the results of my work in one of IJCAI's submission tracks. It is my strong belief that a paper introducing a self-play framework which brings together existing algorithms in the literature under the same roof and also proposes a novel algorithm under the same framework could be accepted in IJCAI. For this I have already assembled a group of 4 researchers which are willing to contribute both theoretically and technically. They are: Sam Devlin (Microsoft Research), Spyros Samothrakis (Lecturer at Essex University), Yuan Gao (PhD student at Upsala University) and myself.

5. **CIG 2019**
   There are no official dates for CIG 2019. Assuming that it follows the same trend as previous years, it will be held in August and the call for papers will end on mid March. If myself or my supervisors think that my work is not in a suitable state to be submited to IJCAI, my goal would shift to submitting it to CIG 2019. I am also open to the idea of submiting to other similar conferences, and I welcome discussions on this topic.

6. **CIG 2019 Fighting Game Competition**
   CIG's competition tracks usually have a deadline in early August. Throughout the first year of my PhD I invested a significant amount of time working on a submission for CIG's Fighting Game Competition. I could use my proposed framework as the theoretical basis for an entry into a competition that I am already familiar with.

4.2 Social Robotics paper

As mentioned in the Context section, I co-authored a paper on social robotics with a research group from Upsala University, Sweeden.

- **ICRA submission**
  ICRA 2019 conference will be held on the 20th of May in Montreal, Canada.

4.3 Mandatory modules

As part of the second year of my PhD, I need to take 30 credits worth of modules. My module choices are

- (20 credits) **Multi-agent Interactions & Games**
  This module is taught during two terms, beginning in October and having one exam in the summer term. This means that during these terms I will need to dedicate a portion of my time to studying the course content and preparing for the Summer exam.
- (10 credits) **Demonstration Development module**
  As part of my optional modules I intend to take the Demonstration Development module. This module will allow me to spend 100 hours of full time job in a project. This project consists on building a device capable of displaying classical arcade games where a learning agent plays the game and improves over time. This device will be framed in a case so that it can be displayed on a wall. The motivation is to have a "moving painting" depicting a game being played with increasing level of proficiency. This project features both hardware and software requirements. Hardware wise, the project will need a raspberry pi, a memory card to host the OS on, a small screen and potentially a custom made case that I could create using the university's 3D printers. A non trivial part of the process would be to assemble all the parts together. Software wise, I need to run the famous Arcade Learning Environment and a machine learning framework to train a playing agent. A task to solve is how to train a computationally expensive deep RL model in a raspberry pi hardware.

There are no explicit ethical issues to be concerned about regarding the next step of my research. My experiments will only involve simulated environments, thus eliminating the potential ethical issues of experimenting with humans, animals or other physical objects. The synthetic simulation data, and datasets generated via simulations will not be of unethical nature. Furthermore, most of these datasets will be destroyed when the simulation where they were created terminates.