Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Kullback-Leibler Policy Chaining for Scalable Lifelong Reinforcement Learning

Christopher Doyle
15323093

February 12, 2019

A Final Year Project submitted in partial fulfilment
of the requirements for the degree of
B.Eng (Computer and Electronic Engineering)

Declaration

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: _____ Date: _____

Abstract

Reinforcement Learning (RL) is the primary learning technique employed by AlphaGo and recent techniques for Lifelong RL (LLRL) have shown great success in getting an RL agent to generalise over a handful of tasks. The problem with an RL agent such as AlphaGo learning multiple tasks however, is that each task's information must be stored in memory, and the more complex the task, the more information that needs to be remembered. We can see how scalability becomes an issue if an AI has to master more than a handful of tasks.

This thesis proposes the Kullback-Leibler Policy Chain approach which aims to help an RL agent to generalise over multiple tasks in a scalable manner. The proposed method of LLRL draws upon techniques from the field of Information Theory that will allow an agent to gauge the importance of memories inherited from previously learned tasks. Inheriting useful information from previously completed tasks could be considered more scalable than storing each and every memory.

Acknowledgements

Thanks Mum!

You should acknowledge any help that you have received (for example from technical staff), or input provided by, for example, a company.

Contents

List of Figures

List of Tables

Nomenclature

A	Area of the wing	m^2
B		
C	Roman letters first, with capitals...	
a	then lower case.	
b		
c		
Γ	Followed by Greek capitals...	
α	then lower case greek symbols.	
β		
ϵ		
TLA	Finally, three letter acronyms and other abbreviations arranged alphabetically	

If a parameter has a typical unit that is used throughout your report, then it should be included here on the right hand side.

If you have a very mathematical report, then you may wish to divide the nomenclature list into functions and variables, and then sub- and super-scripts.

Note that Roman mathematical symbols are typically in a serif font in italics.

1 Introduction

There isn't a human on this earth who can triumph over Deepmind's AlphaGo in the game of Go, however this does not quite make AlphaGo super-intelligent in that a child may beat it at draughts.

University of Oxford philosopher Nick Bostrom defines super-intelligence as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" and we are yet to develop a single AI that can adapt to our varying needs or outperform us across several tasks. Obtaining a more general intelligence requires Lifelong Learning (LL) or more simply, an AI that can remember and learn from all of its life experiences like we do, as opposed to mastering a single environment. These traits are a necessity if we want to create AI's that can master more than a single environment.

This thesis addresses lifelong reinforcement learning (LLRL) which allows a reinforcement learning agent to generalise to more than a single task. Existing "dictionary" techniques store the necessary RL model information for each task individually with a known weakness being scalability. The topic of this thesis, *Kullback-Leibler Policy Chaining* , is the preposed approach of combining existing RL techniques with aspects of Information Theory to achieve LLRL in a more scalable manner.

The motivation behind this thesis is the idea that a RL agent could somehow *remember* the important details from what it has learned in the past, without having to selectively isolate the appropriate information each time. Comparing probability distributions is fundamental to information theory and the divergence (distance) between different distributions will be used to assist our LLRL agent. We can think of the divergence between two different distributions as a difference in knowledge that an agent has between two tasks. The larger the divergence, the less common information that the tasks share. This thesis explores an approach that uses this divergence to isolate the most common pieces of knowledge from past tasks in a way that places phantom memories inside our LLRL agent's brain.

1.1 Model-Based vs Model-Free Learning

Model-based learning is the idea an agent must fully model the world and all of its mechanics. The agent must learn what the world contains, create probability distributions over where actions will lead it, and model how the world will reward it for interacting within its environment. Model-free reinforcement learning requires an agent to learn how to optimally interact with a world in order to maximise rewards which follow strict rules. The primary difference between model-free and model-based learning, is that there is much more to learn in a model-based task. In model-free, the environment will direct the agent's state transitions based on its actions but this is not the case for model-based. In model-based, the agent much also model the probability that actions will result in transitioning to some state, as well as modelling the entire reward system. This thesis focuses not on model-based reinforcement learning, but on the less complex model-free methods.

1.2 Lifelong Learning

Lifelong learning encompasses any field that furthers current machine learning techniques to be sufficiently capable of dealing with real world problems. For example, an AI will have to be able to learn a different task, re-learn an old task, or even re-learn multiple old tasks to be especially useful in the real world. Lifelong learning can be seen to have three primary sub areas of research: Transfer Learning (TL), Multi-task Learning (MTL), and Continuous Learning (CL). The distinction between MTL and CL in a reinforcement learning context is that multitask learning requires a complete change of environment between tasks, whereas CL addresses the re-learning of different tasks within the same environment. This thesis is relevant both the fields of TL and Continuous Reinforcement Learning (CRL) in that an agent will transfer knowledge from previous tasks to assist it the learning of both new and previously learned tasks in the same environment.

1.2.1 Issues with Current CRL Approaches

Many recent CRL approaches adopt a dictionary approach, which involves storing all information known about a learned task. Considerable storage is therefore required if task information contains large multi-dimensional probability distributions over states and their transitions. As required by LL, a CRL agent must be useful in the real world where the number of states can grow exponentially (e.g. if we account for visual and/or

audio input). If a CRL is then required to learn all of the tasks necessary to say, help an elderly person cross the street, then storing all the necessary information for every task begins to become an issue. These scalability restrictions hinder the field of CRL and limit how useful an AI can be in a practical sense. This thesis proposes a new approach that attempts to identify the essential information of a task and to reduce the storage required, while also reducing the time it takes for a CRL agent to learn a previously unseen task. Other issues with current LL techniques are further explored in section XXXXX**2.X** XXXX.

1.3 Thesis Aims and Objectives

The primary aim of this thesis is to deliver a CRL method that can adapt to a continuously changing environment in a scalable manner. The scalability of LLRL techniques is critical if we aim to eventually achieve artificial intelligences that can react and learn when faced with as many data inputs as our own sensory inputs are - without failure. Humans intuitively understand not to walk off a ledge without having personally tried it before, and this is the sense of intuition that I hope to create within an Artificial Intelligence.

1.4 Thesis Assumptions

Not sure what to write here.

1.5 Thesis Contribution

This thesis identifies that the field of LLRL can be complemented by terms originating in Information Theory to create more scalable techniques. Unlike existing LLRL techniques that facilitate continuous learning, the methods outlined in this thesis have minimal storage requirements, and rather focus on creating agents that bequeath their memories of different tasks to create an inherited intuition as it were. The method of chaining together the memories of selective LLRL agents is the primary contribution of this thesis.

1.6 Document Structure

Following on from this introduction chapter, Chapter 2 will explain any required background knowledge on Reinforcement Learning and Information Theory methods. Chapter 3 then presents the proposed experimental theory behind this thesis including any associated reasoning. Chapter 4 will describe the experimental implementation of the theory presented in Chapter 3, and we will then evaluate the results in Chapter 5 with respect to existing LLRL techniques. The thesis concludes with Chapter 6, in which the reader may find a summary of the work and an outline of any remaining or discovered issues.

2 Background and Related Work

This section covers the knowledge necessary to understand how the learning techniques in this thesis, and how information theory can be employed alongside these methods to facilitate continuous reinforcement learning (CRL). First we will look at Reinforcement Learning and the Policy Gradient Approaches that this thesis requires to facilitate CRL. The second second of this chapter explores how exactly information theory can measure the divergence between probability distributions, and how this relates back to CRL.

2.1 Reinforcement Learning

As discussed in Chapter 1, this thesis focusses primarily on model-free reinforcement learning (RL). Reinforcement learning is where an agent must interact with an environment in order to learn how to map situations to actions so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them [1]. The environment in question can be said to be made up of a finite number of states which are described by the set

$$\mathcal{S}$$

.

3 LaTeX

seeing LaTeX, or more properly "LaTeX 2ε", is a very useful document processing program. It is very widely used, widely available, stable and free. Famously, TeX, upon which LaTeX is built, was originally developed by the eminent American mathematician Donald Knuth because he was tired of ugly mathematics books(**?**). Although it has a learning curve (made much less forbidding by online tools and resources – see below), it allows the writer to concentrate more fully on the content, and takes care of most everything else.

While it can be used as a word processor, it is a *typesetting* system, and Knuth's idea was that it could be used to produce beautiful looking books:

> *LaTeX is a macro package which enables authors to typeset and print their work at the highest typographical quality, using a predefined, professional layout.*[1]

LaTeX has great facilities for setting out equations and a powerful and very widely supported bibliographic system called BibTeX, which takes the pain out of referencing.

Three useful online resources make LaTeX much better:

(1) An excellent online LaTeX environment called "Overleaf" is available at `http://www.overleaf.com` that runs in a modern web browser. It's got this template available – search for a TCD template. Overleaf can work in conjunction with Dropbox, Google Drive and, in beta, GitHub.

(2) Google Scholar, at `http://scholar.google.com`, provides BibTeX entries for most of the academic references it finds.

(3) An indispensable and very fine introduction to using LaTeX called *"The not so short introduction to LATEX 2ε"* by **?**) is online at `https://doi.org/10.3929/ethz-a-004398225`. Browse it before you use LaTeX for the first time and read it carefully when you get down to business.

[1] This is from **?**). Did we mention that you should minimise your use of footnotes?

Other tools worth mentioning include:

- `Draw.io` – an online drawing package that can output PDFs to Google Drive – see `https://www.draw.io`.

4 Evaluation

5 Conclusion

A1 Appendix

You may use appendices to include relevant background information, such as calibration certificates, derivations of key equations or presentation of a particular data reduction method. You should not use the appendices to dump large amounts of additional results or data which are not properly discussed. If these results are really relevant, then they should appear in the main body of the report.

A1.1 Appendix numbering

Appendices are numbered sequentially, A1, A2, A3... The sections, figures and tables within appendices are numbered in the same way as in the main text. For example, the first figure in Appendix A1 would be Figure A1.1. Equations continue the numbering from the main text.