Abstract

There isn't a human alive who can triumph over Deepmind's AlphaGo Zero in the game of Go, however this does not make the AI super-intelligent in that a child may beat it at draughts. Obtaining a more general intelligence requires 'lifelong learning', i.e. an AI that can remember and learn from all of its life experiences to master more than a single problem.

This thesis concerns the Kullback-Leibler Policy Chain method which helps an AI to generalise over several tasks in a scalable manner. In reinforcement learning (Sutton & Barto, 1998), the policy is considered to be the instruction set of actions to take given the state of an environment. Existing techniques for continuous reinforcement learning (CRL) such as Composition Value Functions of James et. al (2018) and CAPS of Li et al. (2018) scale poorly as the number of tasks for an agent grow due to the need to store many learned policies. I seek to address this problem by exploiting properties of the Kullback-Leibler Divergence across specific policies in a way that keeps only the most important memories. This results in prior policies for any new task already possessing the most common traits of historic policies.

The need to store historic policies as some existing methods do is greatly reduced with the introduction of a pseudo running total or a *chain* of these polices. Within this policy chain, we can say that the youngest link (i.e. policy for most recent task) will inherit memories from older policies throughout the chain. A further advantage over CAPS is that no source policies need to be provided by a human supervisor.

Introduction

The environment used in this experiment will be a GridWorld of varying sizes. Both the agent's starting location and the goal will be at random locations and the agent will continue to search until the negative reward limit is reached. A Policy Gradient approach will be taken to solve for the optimal policy (see Background). To facilitate an environment of CRL, the goal within the GridWorld will be changed after the policy for that goal is considered optimal. The agent will then be required to learn or re-learn the optimal route to the new goal location from anywhere in the grid. The aim of this experiment is to appropriately select an optimal prior policy for when this occurs, without needing to store a large dictionary of previously learned policies to consult.

Catastrophic forgetting (REF) is an event that can occur when we attempt to solve a multitask environment with traditional policy gradient methods. If we consider each task as a unique goal location within the grid, catastrophic forgetting can take the following form in a RL setting:

1. Optimise policy for task 1.
2. Change location of goal (i.e. begin task 2).
3. Optimise policy for task 2.
4. Agent can no longer perform task 1, and may find it impossible to relearn.
5. Catastrophic forgetting has occurred.

Inspiration for this thesis was originally taken from the research into entropy-regularized policy gradient methods of Mnih et al. (2016) which attempted to address catastrophic forgetting. This experiment regularised the posterior policy of an agent such that the policy π_a of an agent must remain similar to some other policy π_0. The aim of this thesis is to create a cost function that is optimised by the optimal prior policy, such that this optimal policy is highly susceptible to learning useful re-used traits of previous policies. Utilised correctly, a chain of these policies could quickly lead to a prior policy that directly inherits the most common traits of historic policies, without having to store the policies themselves. In the context of the GridWorld environment, it means that the prior policy π_i for goal location i will already know not to leave the grid.

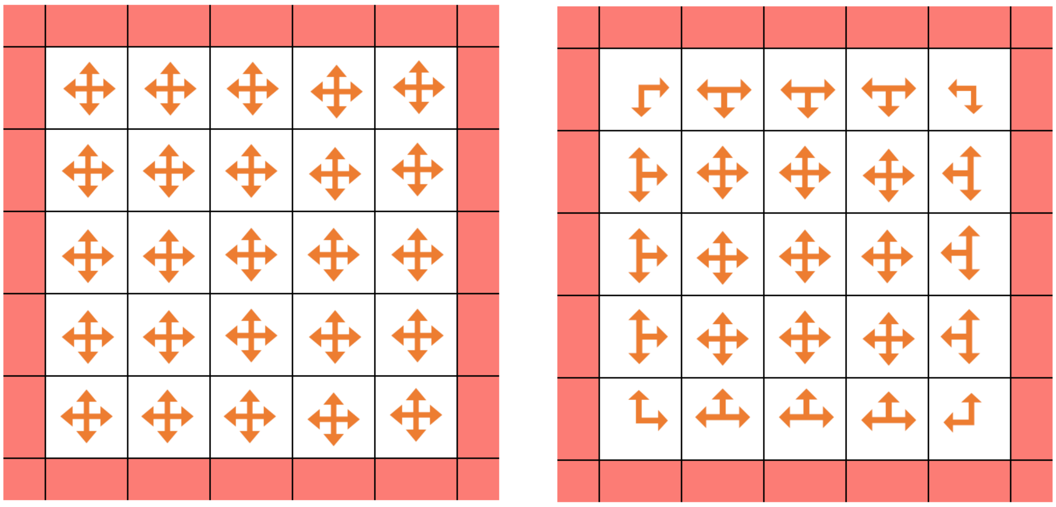

Fig. 1: Uniform prior policy versus worst-case optimal prior policy

It is important to note that the above figures are not truly representative of what occurs in a Kullback-Leibler policy chain (KLPC) and this will be addressed further in the thesis. EDIT: Uniform policies in the inner squares may all have different probabilities of each direction, i.e. P(Left | Inner square) is not necessarily 0.25.

Background

Markov Decision Process

Reinforcement Learning

Policy Gradient Methods

Kullback-Leibler Divergence