Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

Projektarbeit (Informatik)

Reinforcement Learning mit einem Multi-Agenten System für die Planung von Zügen

Autoren	Dano Roost Ralph Meier
Hauptbetreuung	Andreas Weiler
Nebenbetreuung	Thilo Stadelmann

Datum	18.09.2019

Zusammenfassung

Zusammenfassung in Deutsch

Abstract

Abstract in English

(Deutschsprachiges Management Summary)

(Englischsprachiges Management Summary)

Vorwort

Stellt den persönlichen Bezug zur Arbeit dar und spricht Dank aus.

Erklärung betreffend das selbständige Verfassen einer Projektarbeit an der School of Engineering

Mit der Abgabe dieser Projektarbeit versichert der/die Studierende, dass er/sie die Arbeit selbständig und ohne fremde Hilfe verfasst hat. (Bei Gruppenarbeiten gelten die Leistungen der übrigen Gruppenmitglieder nicht als fremde Hilfe.)

Der/die unterzeichnende Studierende erklärt, dass alle zitierten Quellen (auch Internetseiten) im Text oder Anhang korrekt nachgewiesen sind, d.h. dass die Projektarbeit keine Plagiate enthält, also keine Teile, die teilweise oder vollständig aus einem fremden Text oder einer fremden Arbeit unter Vorgabe der eigenen Urheberschaft bzw. ohne Quellenangabe übernommen worden sind.

Bei Verfehlungen aller Art treten die Paragraphen 39 und 40 (Unredlichkeit und Verfahren bei Unredlichkeit) der ZHAW Prüfungsordnung sowie die Bestimmungen der Disziplinarmassnahmen der Hochschulordnung in Kraft.

Ort, Datum:

……

Unterschriften:

……

……..

……

Das Original dieses Formulars ist bei der ZHAW-Version aller abgegebenen Projektarbeiten zu Beginn der Dokumentation nach dem Abstract bzw. dem Management Summary mit Original-Unterschriften und -Datum (keine Kopie) einzufügen.

Contents

1. Einleitung

1.1. Baseline

- Nennt bestehende Arbeiten/Literatur zum Thema -> Literaturrecherche
- Stand der Technik: Bisherige Lösungen des Problems und deren Grenzen
- (Nennt kurz den Industriepartner und/oder weitere Kooperationspartner und dessen/deren Interesse am Thema Fragestellung)

The indirect industry partner during this work was the Swiss Federal Railways (SBB AG) which created the challange on AICrowd[?].
The challenge consists of 2 parts. Part 1 was about avoiding conflicts with multiple trains (agents) on their given environment.
The aim of part 2 was to optimize train traffic which includes trains with different speeds, broken trains and less switchover facilities.
We could use Stefan Husters work as a foundation to build our solution for the challenge.

1.2. Zielsetzung / Aufgabenstellung / Anforderungen

- Formuliert das Ziel der Arbeit
- Verweist auf die offizielle Aufgabenstellung des/der Dozierenden im Anhang
- (Pflichtenheft, Spezifikation)
- (Spezifiziert die Anforderungen an das Resultat der Arbeit)
- (Übersicht über die Arbeit: stellt die folgenden Teile der Arbeit kurz vor)
- (Angaben zum Zielpublikum: nennt das für die Arbeit vorausgesetzte Wissen)
- (Terminologie: Definiert die in der Arbeit verwendeten Begriffe)

The aim of the project was to build and train a model which uses reinforcement learning to optimize train traffic on the flatland simulation.
The work it self consists out of 3 major parts: first round, second round and the attempt to add communication between the agents to prevent them from blocking each other.

2. Technical foundation

2.1. Reinforcement learning

Basic definitions

In recent years, major progress has been achieved in the field of reinforcement learning (RL). In RL, an agent A learns to perform a task by interacting with an environment E. On every discrete timestep t the agent needs to take an action u. The selection of this action u is based on the current observation s. The success of the agent is measured by reward R received. If the agent does well, it receives positive reward from the environment, if it does something bad, there is no or negative reward. The goal of the agent A is now to take an action that maximizes the expected future reward $ER_t\ R_t\ R_t\ s_t$ given the current observation s.
The current observation s_t, also known as the current state is used to determine which action u to take next. An agent can observe its environment either fully or partially.

Value based vs. policy based methods

Reinforcement learning methods are categorized into value-based methods and policy-based methods. Those variants differ on how they select an action u from a state s. Value-based reinforcement learning has its origins in dynamic programming. Through repeated rollouts of the environment, a value function Vs is aquired. Vs aims to estimate the future expected reward for any given state s as precisely as possible. This estimation Vs is achieved by either a lookup table for all possible states or a function approximator. In this work, we solely focus on the case that Vs is implemented in form of a neural network as function approximator. Using this approximation Vs we can now select the action u that takes the agent into the next state s_t with the highest expected reward
The second category of reinfocement learning algrithms are the so called policy based methods. These methods aim to aquire a stochastic policy that maximizes the expected reward R by taking actions with certain probabilities. Taking actions based on probabilities solves an important issue of value based methods, which is, that by taking greedy actions with respect to state s, the agent might not explore the whole state space and misses out on better ways to solve the environment (source!!).

Relation to this work

The goal of this work is to apply an RL algorithm to the vehicle rescheduling problem. Based on the work of S. Hubacher (source!!!), we use a distributed RL algorithm that learns a policy to control the traffic of trains on a rail grid. To do so, we use the asynchronous advantage actor critic algorithm and expand its definiton to the use case of multiple agents.

2.2. The flatland rail environment

The flatland environment is a virtual simulation environment provided by the Swiss Federal Railway SBB and the crowdsourcing platform AICrowd. The goal of this environment is to act as a simplified simulation of real train traffic. Using flatland, we can train RL algrithms to control the actions of trains, based on observations on the grid. Flatland has a discrete structure in both its positions and its

timesteps.The whole rail grid is composed out of squares that can have connections to neighbouring squares. In certain squares, the rails splits into two rails. On those switches, the agent has to make a decision which action it wants to take. Dependent on the type of switch, there are different actions available. All rail parts, independent of if it is a switch also allow to take the actions to do nothing (remain halted, or keep driving), to go forward or to brake. The action space is therefore defined by:

$$U \quad \text{Do nothing, go left, go forward, go right, brake}$$

It is important to note that trains don't have the ability to go backwards and therefore need to plan ahead to avoid getting stuck. To learn which actions to take, the agents have to learn to adapt to an unknown environment due to the fact that the environments are randomly generated and differ on each episode. Depending on the given parameters, the size and complexity of the grid can be adjusted. This allows for dynamically changing the difficulty for the agents.
The goal of each agent is to reach an assigned target train station as fast as possible. Agents that reach this destination are removed from the grid which means, they can no longer obstruct the path of other trains.

Agent evaluation

AICrowd an SBB provide a system for agent evaluation. This system evaluates the policy on a number of unknown environments and outputs the percentage of agents that reached their destination as well as the received reward while doing so. The evaluation reward scheme is thereby as follows:

$$R_t \quad \begin{array}{l} \text{if } s_t \text{ is not terminal} \\ \text{otherwise} \end{array}$$

All submissions to the flatland challenge get graded by the percentage of agents that made it to destionation followed by received reward (IMPROVE, not nicely formulated).

Mnih et al, DQN Atari https://www.cs.toronto.edu/ vmnih/docs/dqn.pdf

Wu et al, A3C https://arxiv.org/abs/1602.01783

Overview over MARL, Hernandez-Leal et al https://arxiv.org/pdf/1810.05587.pdf

A3C in a multi agent environment, https://arxiv.org/pdf/1903.01365.pdf

3. Vorgehen / Methoden

- (Beschreibt die Grundüberlegungen der realisierten Lösung (Konstruktion/Entwurf) und die Realisierung als Simulation, als Prototyp oder als Software-Komponente)
- (Definiert Messgrössen, beschreibt Mess- oder Versuchsaufbau, beschreibt und dokumentiert Durchführung der Messungen/Versuche)
- (Experimente)
- (Lösungsweg)
- (Modell)
- (Tests und Validierung)
- (Theoretische Herleitung der Lösung)

3.1. (Verwendete Software)

Für die vorliegende Arbeit wurden die unten aufgeführten Programme eingesetzt.

Arbeitsumgebung

- Microsoft Windows 8 developer preview

Virtual Machine

- Oracle VM VirtualBox, Version 3.2.10

CAD Catia

- CATIA, Version 5.19 (in VirtualBox)

Dokumentation

- proTeXt mit TexMakerX 2.1 (SVN 1774), latex-project.org
- Microsoft Visio 2007
- Adobe Acrobat 8 Professional 8.1.6

4. Infrastructure

4.1. Infrastructure

We used various computers and servers to train our model. Most of the time was a test environment server of the ZHAW School of Engineering used with 56 CPU cores and 721Gb memory in addition 3 Openstack machines with 8 CPU cores each.
The flatland environment was running on each cpu core and the results where sent to a webserver which updated the model and sent the new neuronal network weights back to the client.
The reason why we used CPU cores over GPU performance is that the reinforcement algorithm A3C which we used performs better on CPUs instead of GPUs.

4.2. Used Software

Towards the end we decided to convert our Python code into C code, to get a performance boost in training our model. Cython was used in this case.

5. Resultate

- (Zusammenfassung der Resultate)

6. Diskussion und Ausblick

- Bespricht die erzielten Ergebnisse bezüglich ihrer Erwartbarkeit, Aussagekraft und Relevanz
- Interpretation und Validierung der Resultate
- Rückblick auf Aufgabenstellung, erreicht bzw. nicht erreicht
- Legt dar, wie an die Resultate (konkret vom Industriepartner oder weiteren Forschungsarbeiten; allgemein) angeschlossen werden kann; legt dar, welche Chancen die Resultate bieten

7. Verzeichnisse

List of Figures

List of Tables

(Glossar)

In diesem Abschnitt werden Abkürzungen und Begriffe kurz erklärt.

Abk	Abkürzung
XY	Ix Ypsilon
YZ	Ypsilon Zet

Listings

A. Anhang

A.1. Projektmanagement

- Offizielle Aufgabenstellung, Projektauftrag
- (Zeitplan)
- (Besprechungsprotokolle oder Journals)

A.2. Weiteres

- CD mit dem vollständigen Bericht als pdf-File inklusive Film- und Fotomaterial
- (Schaltpläne und Ablaufschemata)
- (Spezifikationen u. Datenblätter der verwendeten Messgeräte und/oder Komponenten)
- (Berechnungen, Messwerte, Simulationsresultate)
- (Stoffdaten)
- (Fehlerrechnungen mit Messunsicherheiten)
- (Grafische Darstellungen, Fotos)
- (Datenträger mit weiteren Daten (z.B. Software-Komponenten) inkl. Verzeichnis der auf diesem Datenträger abgelegten Dateien)
- (Softwarecode)