Deep Q-learning Networks in the Gym

HBO-ICT Artificial Intelligence

Studenten: Sinem Ertem & Gijsbert Nutma

Docent en gastdocent: David Isaacs Paternostro & Bas Niesink

Datum: 04-06-2021

Inhoudsopgave

1. Inleiding

Voor het vak ***Adaptive Systems*** zijn we de afgelopen 4 weken bezig geweest om *Reinforcement Learning* naar de praktijk te brengen door middel van het schrijven van een adviesrapport voor een realistische casus. De casus luidt als volgt:

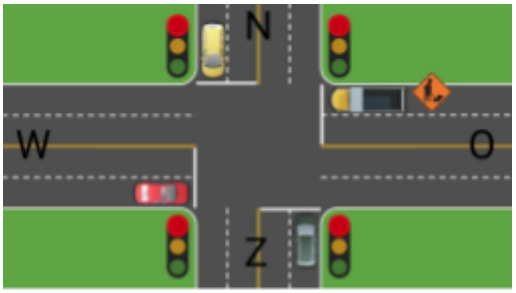
"Sharon Dijksma (burgemeester van Utrecht) wil de doorstroom in Utrecht verbeteren. Hiervoor heeft ze een oproep gedaan aan bedrijven om met innovatieve ideeën te komen. Het bedrijf waar jullie werken ziet hier potentie in en jullie wordt gevraagd te kijken naar mogelijkheden om clusters verkeerslichten in de stad aan de sturen met reinforcement learning."

2. Bestaande oplossingen vs RL

Verkeersregeltechniek houdt in dat verkeerslichten zo optimaal mogelijk worden ontworpen. Verkeerslichten zijn hier samen met o.a. detectielussen onderdeel van. Op dit moment worden er verschillende technieken/algoritmen gebruikt om deze verkeerslichten te besturen. We zullen een paar van deze technieken hier bespreken en afwegen tegen *Reinforcement Learning*.

2.1. Fixed time control

Het *fixed time control* algoritme houdt in dat er een vaste sequence is waar de verkeerslichten zich aan moeten houden. Zo kan het algoritme starten met het meest westelijke verkeerslicht en deze eerst op groen te zetten om het vervolgens na een vast aantal tellen op rood te laat te springen, gevolgd door de zuidelijke op groen te zetten en ook na een vast aantal tellen op rood te laten springen. Dit gaat zo rond tot elk stoplicht zijn beurt heeft gehad en het algoritme weer van begins af aan zijn sequence weer afmaakt.



Figuur 1: Illustratie verkeerskruispunt

2.1.1. Voordeel

Het voordeel van dit algoritme is dat er geen sensor voor nodig is om het algoritme werkend te krijgen. Het gebruikt slechts x ticks als tijd.

2.1.2. Nadeel

Het resultaat van het algoritme is sloom, iedereen moet een vastgestelde hoeveelheid wachten. Doordat het algoritme geen beeld heeft van hoe het verkeer eruit ziet kan er niet aan de hand van de situatie gehandeld worden. Zo is het mogelijk dat er maar 1 auto zich bevindt bij het kruispunt maar nog steeds de gehele sequence moet wachten omdat hij het vorige verkeerslicht gemist heeft.

2.2. Coordinated control

Dit algoritme zorgt er weer voor dat automobilisten een zo lang mogelijk reeks groene verkeerslichten tegenkomen. Deze verkeerslichten zijn aan elkaar gekoppeld en houden de gemiddelde rijsnelheid bij. In Nederlandse termen wordt dit ook wel de *groene golf* (*Willem Wever, 2021*) genoemd. Dit lukt dus alleen als automobilisten zich aan een snelheid houden van x km/uur in de groene golf. Dit heeft als voordeel dat het aantal aanrijdingen wordt verminderd, de wachttijd korter wordt en ook de automobilisten aanmoedigt om binnen de snelheidslimiet te rijden.

2.2.1. Voordeel

Het algoritme heeft controle over meerdere kruispunten met stoplichten voor de gehele groene golf. Voor RL is dit lastig te trainen op grote schaal. Bovendien is het zeer onwaarschijnlijk dat de uiteindelijke ***policy*** globaal optimaal zou zijn zonder de coördinatie tussen lokale leer/optimalisatieprocessen

2.2.2. Nadeel

Door heel Nederland moeten er, waar nodig, groene golf borden /palen worden geplaatst wat extra kosten met zich meebrengt. Verder is een duidelijke doorgaande hoofdstroom vereist om de groene golf te realiseren. Voor veel kruispunten geldt echter dat deze ontbreekt en dat de verkeersvraag min of meer evenredig verdeeld is over de verschillende richtingen.

2.3. Adaptive control

Het *adaptive control* (Federal Highway Administration, z.d.) systeem is een strategie waarin de timing van verkeerslichten veranderen op basis van verkeersaanbod. Als in het oosten meer aanbod is zal het licht daar langer op groen springen in vergelijking met verkeerslichten in andere richtingen waar het aanbod minder is. Hierdoor zorg je er dus voor, in tegenstelling tot *fixed time control*, dat een lange rij met auto's niet hoeven te wachten voor slechts 1 auto. Deze methode heeft altijd een sterkere voorkeur voor doorstroom dan voor eerlijkheid. Dit is vergelijkbaar met een greedy policy als het reward overeenkomt met de lengte van de wachtrijen voor de verkeerslichten.

2.3.1. Voordeel

Dit algoritme houdt real-time rekening met het aantal wachtende voor een verkeerslicht door middel van sensoren. Door het real-time bij te houden werkt het zo optimaal mogelijk.

2.3.2. Nadeel

Er zal altijd een persoon de dupe zijn omdat die gemiddeld iets langer moet wachten dan de rest.

2.4. Algemene voordelen RL

Door camera's in te zetten is de kans dat je de privacy schendt (waarbij ethiek een rol speelt) groter dan wanneer je Reinforcement Learning gebruikt voor verkeersoptimalisatie waarbij je geen camera's nodig hebt. Naast detectielussen worden er ook namelijk camera's (*Verkeer Fandom, z.d.*) ingezet om voertuigen te detecteren en te gebruiken voor het algoritme. Dit is een **voordeel** voor RL**.**

RL zorgt ervoor dat het elke stap vanuit zichzelf leert. Het genereert eigen 'training' ervaringen uit de omgeving. Als voorbeeld onze agent in de *CartPole-v0* game van *openAI Gym.* Het systeem wordt

bestuurd door een kracht van +1 of -1 op het wagentje uit te oefenen. Een beloning van +1 wordt gegeven voor elke tijdstap dat de paal rechtop blijft staan. De episode eindigt wanneer de paal meer dan 15 graden uit het lood staat, of het karretje meer dan 2,4 eenheden van het midden beweegt. Na **649** episodes was het gelukt om constant een reward van 200 te behalen en had de agent eindelijk voldoende kunnen leren door ervaring op de toen in voorgaande episodes. Een atari game is natuurlijk wel lastig te vergelijken met een complexere verkeerssituatie.

Echter is een **nadeel** dat bij supervised learning de vooruitgang bijvoorbeeld vooral te danken is aan grote gelabelde datasets zoals *ImageNet*. Bij RL zou je dus een grote en diverse verzameling van omgevingen nodig hebben. De bestaande open-source collecties van RL-omgevingen zijn echter niet gevarieerd genoeg, en ze zijn vaak zelfs moeilijk op te zetten en te gebruiken.

Nog een groot voordeel van RL is dat een programmeur niet een algoritme hoeft uit te typen tot de kleinste detail. De variabelen kan het RL model zelf bepalen door het gebruik van zijn Q-table.

2.5. Algemene nadelen RL

RL gaat er vanuit dat de wereld waarin we leven Markovian is, dat de acties die genomen worden discreet zijn zoals we in de Jupyter Notebooks hebben gedemonstreerd met de Atari Games en dat de agents in de omgeving Tabula Rasa (*Wikipedia-bijdragers, 2020*) geleerden zijn. Dit is in de psychologie het idee dat de mens geboren wordt zonder enige kennis, vaardigheid en persoonlijkheid. Een **nadeel** is hiervan is dat het niet goed genoeg de realistische wereld nabootst waarin deze kennis en vaardigheden wel belangrijk zijn, al helemaal in het verkeer wat al snel uit de hand kan lopen.

Verder zal het RL langdurig getraind moeten worden en voor een goed werkend systeem zal er zorgvuldig naar de rewards gekeken moeten worden om een werkend model te krijgen.

3. Ethische overwegingen

Artificial Intelligence begint steeds een belangrijkere rol te krijgen in onze maatschappij. Om dit op grote schaal toe te passen is ethiek een gewichtig aspect wat niet vergeten mag worden.

Doordat er jaarlijks nog steeds meer dan 600 verkeersdoden (*Centraal Bureau voor de Statistiek, 2020*) vallen is het van belang dat het gebruiken van nieuwe technieken met behulp van algoritmen zo veilig en zorgvuldig mogelijk gebeurt. Het trainen van het model zal daarom ook in een realistische simulatie moeten plaatsvinden omdat het niet gedaan kan worden op de weg waar automobilisten de dupe van kunnen worden. Dit zou je kunnen zien als een voordeel en nadeel omdat een realistische simulatie niet altijd de werkelijkheid perfect kan nabootsen en daarom de vraag blijft of we het nieuwe getrainde model in de echte wereld kunnen implementeren omdat er jaarlijks nog steeds zo veel verkeersongelukken gebeuren en dit niet verantwoord zou zijn.

Als tweede punt hebben we de bias (reward) die het model meekrijgt om te trainen. Deze bias is van grote invloed op het model omdat dit bepaalt welke actie het model kiest om uit te voeren. Zo zou je een hoge reward kunnen geven als auto's de snelheid kunnen behouden en de wacht-rijen zo kort mogelijk zijn. Ook is het mogelijk om de flow van het verkeer te bepalen door de bias aan te passen. Zo kun je de reward verhogen voor auto's die een bepaalde richting opgaan.

4. Praktische overwegingen

In dit hoofdstuk behandelen we kort wat de praktische overwegingen zijn en in hoeverre het uitvoerbaar is in de praktijk.

Als eerst hebben we cluster verkeerslichten die moeten bestaan uit allemaal opeenvolgende verkeerslichten. Hierbij is het cruciaal dat elk van deze verkeerslichten verbonden zijn aan het algoritmen. Omdat dit nog niet door heel Nederland zo is geregeld met de verkeerslichten, zal dit extra veel tijd vergen. In kleinere schaal zal het makkelijker uitvoerbaar zijn.

Ten tweede hebben we een belangrijk voorwaarde en dat is de hoeveelheid data. Om het algoritme meer data te geven kun je meerdere lussen in de weg te leggen. Hierdoor kan de snelheid en hoeveelheid van auto's op meerdere punten op de weg gemeten worden. Meer data kan zorgen voor een accuraat RL model van de agent.

Ten derde hebben we de rewards. Bij RL moeten de rewards aangegeven worden en het aanmaken van deze rewards zijn de kern waarop het algoritme te werk gaat. Deze rewards zullen zorgvuldig aangemaakt moeten worden en ook rekening gehouden moeten worden met welke combinatie tussen verkeerslichten een botsing plaats kan vinden om deze te blokkeren door middel van een negatieve reward als het algoritme deze combinatie van verkeerslichten uitvoert.

Als laatst hebben we de training simulatie. Deze zal veel moeten lijken op de real world scenario. zodat het model accuraat getraind kan worden. Hierbij is van belang dat de detectie methode goed gesimuleerd wordt. Een voordeel hiervan is dat uitgezocht kan worden of de eerdere vraag naar meerdere lussen impact zal hebben en waar deze lussen zich zouden moeten bevinden door ze op meerdere plekken tussen de kruispunten neer te leggen.

5. Globale planning

Alles wat we in de vorige hoofdstuk hebben besproken moet natuurlijk ook gepland worden, mocht het geïmplementeerd worden. Daarom bespreken we in dit hoofdstuk hoe de globale planning eruit ziet, hoeveel tijd dit kost en wat de risico's zijn (kans en impact).

- Voor de veiligheid van automobilisten zal het RL algoritme getraind moeten worden in een simulatie omgeving. Er zijn al wel bekende simulatie programma's voor het testen van modellen waaronder *SUMO* (*Eclipse, z.d.*). Het is van groot belang dat deze omgeving realistisch data genereert voor het model. De juiste data heeft de grootste impact op het model daarom is het opstellen van de simulatie de eerste en een van de grootste taken. Er zal hier flink wat tijd in zitten om het data zo accuraat mogelijk te maken en zal naar geschatte tijd wel een gehele sprint (twee weken) kunnen duren. Het risico van deze taak is dat de data niet lijkt op de real world scenario, wat tot gevolg kan hebben dat het model niet goed genoeg of zelfs niet werkt.

- De rewards zullen vastgesteld moeten worden voordat er getraind kan worden op het model. Hier kan over gediscussieerd worden wat de reward zal zijn voor het model. Denk hierbij aan de lengte van de wachtrijen, de doorstroom van auto's, de duratie dat automobilisten moeten wachten en bijvoorbeeld of het rijden buiten de stad een hogere reward krijgt om zo de uitstroom te vergroten vergeleken met de instroom richting het centrum. Ook zal er een rule-based systeem moeten worden gecreëerd om alle gevaarlijke keuzes van het model te verbieden, om zo de mogelijkheid van botsingen door slechte light signalen te voorkomen. Naar verwachting zal dit maximaal een week hoeven te duren. Het risico is dat een reward te hoog is waardoor het model hier alleen maar op focused. De rewards zullen daarom goed uitgedacht moeten worden.

- Als de vorige punten behandeld zijn kan het RL model opgesteld worden en getraind worden. Het opstellen zal hierbij naar ervaring het makkelijkst zijn het het trainen wat frustrerend zijn. Omdat het uitvogelen van alle parameters lang kan duren. De duratie van deze taak is slecht te voorspellen maar naar verwachting zal na een gehele sprint wel resultaat moeten zijn. Het risico is dat het model niet configureert.

- Daarna kan het model als het getraind is, gevalideerd worden. Door zorgvuldig het model uit te lezen.

- Als het model gevalideerd is kan het geïmplementeerd worden in de echte wereld. De duratie is voornamelijk afhankelijk van het technische team van de verkeerslichten.

- In een test periode moet het model gemonitord worden om zeker te zijn of het model het goed werkt op de kruispunten. Zo niet moet het d.m.v. een noodknop, mogelijk zijn om terug te gaan naar het oude algoritme.

- Als het model door de testperiode heen komt is het goedgekeurd en zou eventueel verdere verbeteringen kunnen worden uitgevoerd. Het zou interessant zijn om het model verder te trainen met de echte data om zo wellicht meer rekening te houden met de verkeersstroming.

6. Conclusie

In de vier weken dat we les hebben gekregen over *Reinforcement Learning* hebben we vooral mogen werken aan de twee notebooks. De eerste notebook ging voornamelijk in op *Q-Learning* met behulp van *Q-tables*. RL zou een goede oplossing kunnen zijn om verkeerslichten te sturen. Uit de ervaring die we hebben opgedaan blijkt namelijk dat je door het implementeren RL aan de hand van *Q-learning* een goed werkend model kan maken. Je kunt gewenste keuzes maken op basis van de rewards van de acties.

Het werkelijk implementeren en trainen van het *RL* model zal natuurlijk wat lastig worden. Bij RL moeten de rewards aangegeven worden en het aanmaken van deze rewards zijn de kern waarop het algoritme te werk gaat. Deze rewards zullen zorgvuldig aangemaakt moeten worden en ook rekening gehouden moeten worden met welke combinatie tussen verkeerslichten een botsing plaats kan vinden om deze te blokkeren door middel van een negatieve reward als het algoritme deze combinatie van verkeerslichten uitvoert.

Verder is het lastig om een globale *policy* grootschalig toe te passen op meerdere kruispunten waar de situaties anders kunnen zijn. Echter als het allemaal goed verloopt denken wij dat het regelen van echt verkeer over de cluster verkeerskruispunten beter zal verlopen. Voor de veiligheid van automobilisten zal het RL algoritme eerst getraind moeten worden in een simulatie omgeving zoals in *SUMO*.

Ethisch kan het nog een lastig punt zijn omdat het testen van het *RL* model in een realistische simulatie gebeurt maar dit nog steeds lastig na te bootsen is in de werkelijkheid waar je het test op echte mensen. Het feit dat er al jaarlijks honderden mensen komen te overlijden maakt dit lastiger en je draagt hierbij een groot verantwoordelijkheid. Verder is de bias (reward) van grote invloed op het model omdat dit bepaalt welke actie het model kiest om uit te voeren. Zo kun je een hoger reward geven aan auto’s die zich aan de snelheid houden.

7. Bronnenlijst

Centraal Bureau voor de Statistiek. (2020, 15 oktober). Hoeveel mensen komen om in het verkeer? https://www.cbs.nl/nl-nl/visualisaties/verkeer-en-vervoer/verkeer/hoeveel-mensen-komen-om-in-het-verkeer-

Eclipse. (z.d.). SUMO. Geraadpleegd op 4 juni 2021, van https://www.eclipse.org/sumo/

Federal Highway Administration. (z.d.). EDC-1: Adaptive Signal Control Technology | Federal Highway Administration. Fhwa. Geraadpleegd op 3 juni 2021, van https://www.fhwa.dot.gov/innovation/everydaycounts/edc-1/asct.cfm

Hoe werkt een groene golf in het verkeer? (2021, 23 februari). Willem Wever. https://willemwever.kro-ncrv.nl/vraag_antwoord/de-maatschappij/hoe-werkt-een-groene-golf-het-verkeer

Wikipedia-bijdragers. (2020, 27 januari). Tabula rasa (psychologie). Wikipedia. https://nl.wikipedia.org/wiki/Tabula_rasa_(psychologie)

Verkeer Fandom. (z.d.). Verkeersmonitoring met detectiecamera's | Verkeersmanagement | Fandom. Verkeersmanagement. Geraadpleegd op 3 juni 2021, van https://verkeer.fandom.com/wiki/Verkeersmonitoring_met_detectiecamera%27s