Deep Q-learning Networks in the Gym

Adviesrapport

Casper W. Smet (1740426) & Thijs van den Berg (1740697)

1 Inleiding

In dit adviesrapport wordt antwoord gegeven aan de oproep van Burgemeester Sharon Dijksma van de gemeente Utrecht. De mogelijkheden omtrent het gebruik van Reinforcement Learning (RL) voor het verbeteren van de verkeersstromen zijn in dit adviesrapport beschreven.

Voor deze opdracht kijken wij naar het besturen van alle circa 200 verkeerslichten in Utrecht (Gemeente Utrecht, n/a) door middel van reinforcement learning. Wij nemen alle verkeerslichten mee in deze casus om te vermijden dat problemen qua doorstroom niet simpelweg verplaatst worden naar een ander verkeerspunt.

2 Reinforcement learning en bestaande oplossingen

In dit hoofdstuk worden verschillende voor- en nadelen beschreven van het gebruik van Reinforcement Learning (RL) technieken ten opzichte van andere oplossingen.

2.1 Garantie optimale oplossing

Bij RL is het niet gegarandeerd dat je op de optimale oplossing komt. Je kan een lokaal optimum vinden, of zelfs het probleem oplossen op de *verkeerde* manier. Een goed voorbeeld van dit laatste gebeurde bij de OpenAI Gym *Hide and Seek* environment (Baker et al., 2019). Deze environment bestaat uit twee teams van agents, de *Hiders* en de *Seekers*.

Een van de strategieën van de *Hiders*, aangeleerd door middel van RL, was het vastzetten van de *Seekers* zodat zij niet meer konden bewegen.

Mensen denken uit een bepaalde context, uit de naam “Hide and Seek” begrijpen zij bijvoorbeeld al dat het hun doel is om te verstoppen. De enige context waarvan een RL-algoritme zich bewust is komt van de mens. De mens stelt: het is jouw doel om zo lang mogelijk niet gezien te worden.

2.2 Simulatie vertalen naar de realiteit

RL-algoritmen hebben in het algemeen veel data nodig om te leren. De data kan verzameld worden door gebruik te maken van reeds bestaande clusters van verkeerslichten. Mocht de kwantiteit en kwaliteit van de data niet voldoende toereikend zijn dan kan er uitgeweken worden naar het gebruik van simulaties.

Door middel van simulaties kan er op een grotere schaal data worden verzameld dan in de praktijk. Tevens kan een prototype-versie van het RL-algoritmen niet in de praktijk worden gebruikt om verdere data te verzamelen en resultaten te genereren, dit kan zeer onveilige situaties opleveren.

Het nadeel van het gebruik van simulaties voor de dataverzameling is dat veelal de randgevallen pas in de praktijk naar voren komen. Er dient goed gemonitord te worden of de resultaten die in een simulatie worden behaalt zich ook vertalen naar de werkelijkheid. Idealiter wordt er een zogeheten ‘digital twin’ van de stad ontwikkeld.

Een digital twin is een digitale kopie van bijvoorbeeld een watercentrale, vliegtuig of in dit geval een stad. Verschillende lokale overheden en bedrijven zoals Google en IBM zijn al aan de slag met digital twins (Bliss, 2019).

Het laten draaien van een simulatie en het trainen van een RL-algoritme is echter wel een zeer computationeel intensief proces. Zoals vernomen tijdens de OpenAI-programmeerweken is toegang hebben tot een sterke compute cluster geen overbodige luxe.

2.3 Onvoorspelbaarheid en explainability van AI

De keuzes van AI-oplossingen, met name zij die gebruik maken neurale netwerken, zijn soms moeilijk uitlegbaar. Als wij als mens de besluitvorming van het algoritme niet begrijpen, kunnen wij ook niet (voor ons) onverwachte beslissingen voorzien en tegenhouden. Voor een casus zoals deze waarbij mensenlevens zijn gemoeid is het van belang dat de beslissingen van het algoritme duidelijk zijn.

Het huidige systeem is vermoedelijk gebaseerd op een vaste cyclus van stappen. De programmeur én de eindgebruiker weten beiden wat zij te wachten staat. In dit geval is duidelijk welke beslissingen er genomen worden en kunnen onveilige beslissingen worden tegengegaan.

3 Ethische overwegingen

In dit hoofdstuk worden verschillende ethische overwegingen van de casus besproken.

3.1 Voorrangsregels 2.0

Middels de reward-functie zal het RL-systeem kiezen welke weggebruikers op welk moment in beweging mogen komen. Hierbij moeten verschillende belangen in consideratie worden genomen.

Enerzijds speelt de doorstroming mee, hoe meer weggebruikers de cluster per uur kan afhandelen, hoe beter. Anderzijds moet de wachttijd van de individu ook meegenomen worden. Het kan niet zo zijn dat de doorgaande weg continue op groen blijft waardoor oma tot na de spits moet wachten om over te kunnen steken.

3.2 Korte en lange termijn

Alsmaar de auto's voor laten gaan voor de doorstroming zal op de korte termijn de CO_2-uitstoot doen verminderen, de auto's hoeven niet meer te stoppen en vervolgens accelereren. Fietsers zijn hiervan de dupe, zij stoten geen CO_2 uit dus kunnen zij gerust stil staan voor het stoplicht. Op de lange termijn motiveert dit systeem om met de auto te gaan.

Als het systeem ervoor kiest om een groter belang te hechten aan voetgangers en fietsers dan zal dat op de korte termijn meer uitstoot opleveren maar mogelijk op de lange termijn motiveren om een zuiniger alternatief dan de auto te kiezen.

3.3 Dataverzameling

Om het RL-model te trainen is er veel data nodig. Een deel van deze data kan in simulaties worden gegenereerd. Doch voor de validatie zou het beter zijn om van 'echte' data gebruik te maken. De te verzamelen data kan privacygevoelig zijn. Indien er bijvoorbeeld camerabeelden gebruikt worden kunnen kentekens en gezichten van mensen in beeld komen. Er moet nagedacht worden over manieren om de data te anonimiseren en of sommige data wel überhaupt verzameld dient te worden. Te allen tijde dient de data getoetst te worden aan de geldende data en privacy-wetgeving waaronder de AVG.

3.4 Open source

Het RL-systeem zal ontwikkeld worden met publiek geld. De overheid streeft ernaar om software dat met publiek geld is ontwikkeld, zo veel mogelijk open source uit te brengen (Ministerie BZK, 2021). Onder de voordelen noemt de overheid onder andere transparantie en informatieveiligheid. Het hebben van een extra paar ogen kan zorgen voor een veiliger systeem. Onder andere bugs kunnen eerder opgemerkt worden. In het geval van code-contributies dient men wel waakzaam te blijven en te kunnen garanderen dat door de contributies geen nieuwe bugs of onveilige situaties ontstaan.

Mochten er ongelukken gebeuren als gevolg van code-contributies dan moet duidelijk zijn wie de eindverantwoordelijke was. Is dat degene die de contributie doet, de eigenaar van de code of is dat degene die de contributie heeft toegestaan? Afspraken hierover dienen voor een ieder duidelijk te zijn.

4 Praktische overwegingen

In dit hoofdstuk worden verschillende praktische overwegingen besproken.

4.1 Schaalbaarheid

AI-systemen zijn goed in het vinden van patronen in grote hoeveelheden data. Het algoritme zal naarmate er meer en grotere clusters komen een steeds groter voordeel krijgen ten opzichte tot de mens.

Om het algoritme tot het volledige potentieel te benutten dienen er zo veel mogelijk clusters te worden aangesloten aan het algoritme. Indien het systeem crasht, dienen individuele clusters nog wel correct en met name veilig te werken.

4.2 Vierseizoenen Oplossing

Auto’s hebben in de winter een langere remweg en rijden auto’s gemiddeld slomer. Met dezelfde timing is de doorstroom in de winter dus lager dan in de zomer. Het systeem moet zichzelf af kunnen stemmen op de verschillende weersomstandigheden. Hier moet tijdens het trainen al rekening mee worden gehouden om bias tegen te gaan.

5 Planning

De globale planning kan er als volgt uitzien:

- Functionele en niet-functionele requirements opstellen met verschillende stakeholders
    - Stakeholders waaronder de Gemeente Utrecht, omwonende, uitvoerder
    - Requirements kunnen bijv. opgesteld worden volgens ISO 25010
- Data verzamelen
    - Gebruikmakende van bestaande stoplichten en clusters
    - Uitbreiden met simulatie-data
    - Kwalitatief onderzoek naar weggebruikers en omwonenden
        - Huidige situatie in kaart brengen
- Environment bouwen
- RL-agent schrijven
- RL-agent trainen in environment
- RL-agent testen in environment
- Agent goed laten keuren voor praktijktesten
    - Ethische toetscommissie
    - Toetsen aan de AVG-wetgeving
- Agent uitrollen voor praktijktesten
    - Beginnen bij een enkele cluster en langzaam uitbreiden
- Iteratief de agent verbeteren
- Definitieve toetsing
- Definitieve uitrol naar praktijk
    - Continu het systeem monitoren
    - Agent verder trainen indien nieuwe situaties zich voordoen
- Kwalitatief onderzoek naar weggebruikers en omwonenden
    - Nieuwe situatie vergelijken met oude situatie

Uit onze ervaring met de OpenAI-programmeerweken blijkt dat het optimaliseren van de RL-agent een zeer tijdsintensief proces is. De tijdsintensiviteit van dit probleem is echter te minimaliseren door te investeren in een krachtige compute cluster.

6 Conclusie

Er zijn ongeveer 200 verkeerslichten in Utrecht (Gemeente Utrecht, n/a). Ieder verkeerslicht moet individueel aangestuurd worden. Om hier één gecoördineerd, door RL aangestuurde netwerk van te maken is een gigantische klus.

Ten eerste is er een grote hoeveelheid data nodig om een RL-model te trainen. Het verzamelen hiervan uit de echte wereld is een tijd- en kostenintensief proces. Verder is het mogelijk om een verkeerssimulatie op te zetten voor de stad Utrecht. Hieruit kan synthetische data verzameld worden, welke de “echte” data aanvult.

Ten tweede moet een RL-model afgestemd worden op het scenario; de hyperparameters moeten worden geoptimaliseerd, en belangrijker nog, de rewardfunctie moet gedefinieerd worden. De hyperparameter optimalisatie wordt gedaan door een AI-expert. Het definiëren van de reward is echter nog ingewikkelder.

Het instellen van de rewardfunctie benodigd zowel kennis over RL en het domein. Oftewel, er is nog steeds een verkeersexpert nodig. Zelfs met de betrekking van een domeinexpert is het afstemmen van een RL-model een ingewikkeld en iteratief (lees tijdsintensief) proces.

Dit brengt ons echter bij een tweede probleem met RL. Het model lost het probleem vaker niet dan wel op een gewenste manier op. In deze casus zou bijvoorbeeld alle stoplichten op groen zetten een oplossing zijn. Dit zou echter kunnen resulteren in meer ongelukken.

Een RL-model is niet op dezelfde manier bewust van de context van het probleem als de mens. Het enige dat het model weet, is dat bij actie A de reward omhoog gaat. Dat is verder dus ook het enige waarop het model leert.

Het scenario dat hierboven genoemd wordt is redelijk simpel te vinden door middel van het testen van het model. Het is echter heel goed mogelijk dat andere gevaarlijke keuzes alleen maar voorkomen in niché scenario’s. De kans dat al deze scenario's afgevangen kunnen worden tijdens het maken van het model achten wij nihil.

Verder brengt dit ons tot een belangrijk ethisch vraagstuk: wie is verantwoordelijk wanneer er iets fout gaat? De ontwikkelaars, verkeersexperts, wetgevers of een andere participant in het ontwikkelingsproces?

Op basis van deze argumentatie zijn we tot de volgende conclusie gekomen:

Een reinforcement learning model directe controle geven over de verkeerslichten van Utrecht is onverantwoordelijk.

Dit betekend echter niet dat er helemaal geen plaats is voor het ontwikkelen van een RL-model voor deze casus. In plaats van dat het model directe controle heeft, kan de output van het model gecontroleerd worden door een op-regels-gebaseerd systeem. Dit systeem waarborgd dan de veiligheid.

Een tweede optie is om het model überhaupt niet direct te betrekken bij het besturen van verkeerslichten. Het namelijk ook mogelijk om door middel van observatie van het model in de simulatie regels op te stellen. Deze regels zouden dan direct geïmplementeerd kunnen worden in een op-regels-gebaseerd systeem. Op deze manier is het uiteindelijke besturingssysteem van verkeerslichten simpeler, en dus minder kwetsbaar.

7 Bronnen

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). *Emergent Tool Use from Multi-Agent Interaction* (A. Pilipiszyn, Ed.). OpenAI. https://openai.com/blog/emergent-tool-use/

Bliss, L. (2019, 7 19). Why Real-Time Traffic Control Has Mobility Experts Spooked. *Bloomberg*. https://www.bloomberg.com/news/articles/2019-07-19/why-cities-want-digital-twins-to-manage-traffic

Gemeente Utrecht. (n/a, n/a n/a). *Oproep verkeerslichten*. Gemeente Utrecht. https://www.utrecht.nl/wonen-en-leven/verkeer/verkeersprojecten/verkeerslichten

Ministerie BZK. (2021, 7 1). *Open Source*. Digitale Overheid. Retrieved 6 1, 2021, from https://www.digitaleoverheid.nl/dossiers/oss-kennisnetwerk/