Control of complex actions in humans and robots

Kleijn, R.E. de

Summary in Dutch
Nederlandse samenvatting

Robots en kunstmatige intelligentie zijn de afgelopen jaren een steeds belangrijkere rol gaan spelen in zowel ons alledaagse leven (bijvoorbeeld zelfrijdende auto's) als in het bedrijfsleven. Maar naarmate de taken die robots moeten uitvoeren complexer worden, en uiteindelijk zelfs alledaagse, menselijke taken zouden moeten overnemen, wordt de aansturing van deze robots steeds lastiger.

Hoewel alledaagse taken zoals koken en stofzuigen op het eerste gezicht niet erg ingewikkeld klinken, blijkt het behoorlijk uitdagend om robots te ontwikkelen die deze taken succesvol kunnen uitvoeren. Het ontwikkelen van zo'n robot was het doel van het *RoboHow*-project. Het proefschrift dat voor u ligt is het resultaat van dit project, uitgevoerd door een internationaal consortium van robotici, computerwetenschappers en cognitief psychologen verbonden aan vijf universiteiten en twee onderzoeksinstituten verspreid over Europa.

De wisselwerking tussen cognitief-psychologisch onderzoek en kunstmatige intelligentie werd goed duidelijk tijdens de *neo-cognitieve revolutie*. Behaviorisme, het standpunt dat alleen meetbare en observeerbare gedragingen van mens en dier onderwerp zouden moeten zijn van onderzoek, bleek niet houdbaar vanwege de beperkte verklaringscapaciteit. Typisch "menselijke" fenomenen zoals taal en geheugen konden vanuit een puur behavioristisch framework niet onderzocht worden; zij vereisen een onobserveerbare mentale toestand. De neo-cognitieve revolutie opende de deur voor theorieën die deze mentale toestand—of cognitieve

processen—probeerden te beschrijven en te begrijpen. Maar hoe wordt die mentale inhoud nu precies gerepresenteerd?

Cognitie in mensen en robots

In de kunstmatige intelligentie wordt er in deze gevallen vaak gebruikt gemaakt van een *planner*. In planners worden subacties vaak symbolisch gerepresenteerd als losse eenheden waarop bewerkingen kunnen worden toegepast, en zo kan worden berekend welke subacties er in welke volgorde nodig zijn om vanuit een beginpositie een eindpositie te bereiken. Bij mensen werd lange tijd gedacht dat een soort actieketting verantwoordelijk was voor het uitvoeren van zulke sequentiële actie; de zintuiglijke waarneming van de effecten van een actie zouden functioneren als een *trigger* voor het uitvoeren van de volgende actie. Halverwege de twintigste eeuw werd echter duidelijk dat deze theorieën niet correct konden zijn, onder meer omdat bleek dat sequentiële actie ook kan worden uitgevoerd door mensen waarbij zintuiglijke terugkoppeling verstoord is. Naarmate het bewijs tegen deze symbolische theorieën zich opstapelde werd duidelijk dat het subsymbolische, sensorimotorische aspect van motoracties essentieel was voor sequentiële actieplanning.

Na het *plannen* van een actie moet deze ook *uitgevoerd* worden door het motorsysteem. Een robot die aan de lopende band werkt kan voorgeprogrammeerde acties uitvoeren met een *feedforward*-systeem, waarin informatie uit het programma (bijvoorbeeld "roteer motor 12 naar positie 82,5°") direct wordt omgezet in een motorbeweging. Hoewel dit een zeer snelle manier van aansturing is, kan dit echter voor problemen zorgen in minder voorspelbare omgevingen: wanneer het te manipuleren object zich in positie 83,5° bevindt zal de actie mogelijk mislukken. Een *feedback*-systeem gebruikt informatie uit de omgeving om de actie moduleren. Dit vergroot de kans op een succesvol uitgevoerde actie, maar afhankelijk van de snelheid van de feedback-loop zal de uitvoer minder snel zijn. Menselijk gedrag is het product van een hybride feedforward-feedback-systeem, waarbij een feedforward actieplan wordt gegenereerd

waarin onbekende parameters online kunnen worden ingevuld door een feedback-mechanisme.

Het leren van sequentiële actie

Een manier om deze actieplanning en -uitvoer te onderzoeken is de serial response time (SRT)-taak, geïntroduceerd door Nissen & Bullemer [107]. In deze taak zit de proefpersoon tegenover een beeldscherm waar aan de onderkant een visuele stimulus verschijnt op één van vier mogelijke posities. Wanneer een stimulus verschijnt, drukt de proefpersoon zo snel mogelijk op een knop die onder deze stimulus is gepositioneerd. De proefpersonen weten niet dat de stimuli verschijnen in een vaste, herhalende volgorde. Hoewel deze taak veelvuldig in de literatuur is gebruikt, heeft het als groot nadeel dat de informatie die verzameld wordt gelimiteerd is door de discrete vorm van de respons. Hierdoor is het niet mogelijk om informatie te verzamelen over processen die actief zijn tijdens het inter-trial interval (ITI), zoals voorspellende bewegingen.

In hoofdstukken 4 en 5 worden studies beschreven die een continue variant van de SRT-taak gebruiken, waarbij de vier stimuli en knoppen zijn omgezet in vier zwarte vierkanten op een beeldscherm. Analoog aan de originele SRT-taak worden proefpersonen gevraagd om zo snel mogelijk te reageren op een oplichtende stimulus (het target) door de muiscursor erheen te bewegen. In deze studie werden twee condities gebruikt: een deterministische conditie waarin de stimuli oplichtten in een vaste, herhalende reeks van 10 targets, en een conditie waarin de volgorde van targets willekeurig werd bepaald. In totaal werden 800 targets gepresenteerd. Deze variant van de SRT-taak produceerde dezelfde effecten als de originele taak, waarin proefpersonen sneller worden naarmate het experiment vorderde, maar dit effect was sterker voor de deterministische conditie. Dit duidt op het impliciet leren van de reeks in de deterministische groep, een conclusie eerder getrokken door Nissen & Bullemer [107]. Dankzij de continue aard van de adaptatie die in onze studie is gebruikt, kon duidelijk worden gemaakt dat deze versnelling *niet* kwam

door een simpele versnelling van de motoractie, maar (mede) werd veroorzaakt door voorspellende bewegingen richting de volgende stimulus tijdens het ITI.

Ook werd duidelijk dat proefpersonen twee strategieën kunnen gebruiken om zo snel mogelijk te reageren. Eén groep proefpersonen maakte actief een voorspelling van de volgende target, en bewoog de muiscursor al voordat de target zichtbaar werd in de juiste richting. De tweede groep proefpersonen bewoog de muiscursor naar het midden van het scherm, op gelijke afstand van alle stimuli. Dit is een optimale positie als er geen voorspelling kan worden gemaakt van de volgende target. Dit bleek ook bij het uitvragen van de reeks na afloop van het experiment: proefpersonen die de reeks expliciet hadden geleerd maakten meer voorspellende bewegingen, proefpersonen die de reeks niet expliciet hadden geleerd waren meer geneigd om de muiscursor naar het midden van het scherm te bewegen. Dit laat zien dat mensen—onafhankelijk van hun kennis—een strategie hanteren die optimaal is gegeven hun kennis.

Het modelleren van reinforcement learning

De SRT-taak heeft een beperkte ecologische validiteit, omdat mensen in het dagelijks leven niet simpelweg reageren op stimuli, maar hun omgeving exploreren en leren van interactie met objecten. Om deze reden hebben we een tweede variant van de SRT-taak gemaakt, gebruikmakend van een *reinforcement learning*-paradigma. In deze taak werd niet langer gereageerd op één van de vier oplichtende stimuli, maar moesten de verschillende alternatieven worden uitgeprobeerd waarna feedback werd gegeven over de correctheid van de keuze. Op deze manier werd dezelfde reeks als in de deterministische conditie van de SRT-taak afgewerkt, opnieuw samengesteld uit een 80 maal herhalende reeks van lengte 10. Voor iedere correcte beweging verdiende de proefpersoon 1 punt, voor iedere foutieve beweging verloor de proefpersoon 1 punt. Er bleek verrassend veel variatie te zitten in het aantal behaalde punten na het voltooien van 800 correcte bewegingen.

Reinforcement learning is een techniek uit machine learning die kan leren welke actie moet worden genomen in welke staat om een beloning te maximaliseren, geïnspireerd door operante conditionering. Reinforcement learning-modellen onderscheiden zich van andere technieken zoals supervised learning doordat zij geen gebruik maken van gelabelde trainingsdata, maar door een proces van trial-and-error leren welke acties de meeste beloning opleveren. Dit doen zij door een verwachte beloning, een *Q-value*, toe te kennen aan combinaties van *states* en *actions*.

Om het gedrag van proefpersonen beter te onderzoeken hebben we geprobeerd drie bestaande reinforcement learning-modellen toe te passen op de verzamelde data: (1) Q-learning, (2) SARSA, en (3) Q-learning met eligibility traces. Geen van de onderzochte modellen kon de hoogste scores van proefpersonen evenaren. Dit heeft vermoedelijk te maken met de gebruikte actie-selectiestrategie. Bij het gebruik van een *softmax* actieselectiestrategie zouden mogelijk betere resultaten kunnen worden verkregen, dit is onderwerp van vervolgonderzoek.

Inter-individuele verschillen in prestatie

Hierna hebben we een grotere groep proefpersonen getest, en hebben we een aantal additionele taken afgenomen om te onderzoeken of de verschillen tussen proefpersonen werden veroorzaakt door het al dan niet moedwillig kiezen van verschillende strategieën of beperkingen in werkgeheugen of IQ. Uit eerder onderzoek is bekend dat proefpersonen onder sommige omstandigheden een strategische keuze kunnen maken tussen verschillende manieren van handelen [160]. In *stimulus-based control* delegeert de proefpersoon controle aan de externe stimulus. Versnelling zal hier veroorzaakt worden door het versneld reageren op de stimulus. In *plan-based control* maakt de proefpersoon een interne representatie van een motorplan. Hier kan de proefpersoon actief een voorspelling maken van de volgende stimulus.

In de SRT-taak werd leerprestatie, gemeten door de hoeveelheid expliciete kennis van de reeks, voorspeld door de capaciteit van het visuospatieel

werkgeheugen. In de reinforcement learning-taak werd prestatie, gemeten door de totale score aan het einde van de taak, voorspeld door zowel IQ als de capaciteit van het visuospatieel werkgeheugen. Dit suggereert dat de verschillen in prestatie niet veroorzaakt werden door het kiezen van verschillende strategieën, maar door cognitieve beperkingen.

Het modelleren van optimale bewegingen

Om de geoptimaliseerde muisbewegingen die zichtbaar waren in hoofdstukken 4 en 5 nader te onderzoeken, hebben we in hoofdstuk 6 een robotarm gesimuleerd, aangestuurd door een kunstmatig neuraal netwerk. Deze robotarm kreeg dezelfde SRT-taak als proefpersonen, en de netwerken werden met een evolutionair algoritme getraind om zo snel mogelijk de stimulus die actief werd aan te raken. Er werden drie condities gebruikt: (1) een conditie waarin het netwerk nauwkeurige informatie kreeg over de volgende target (perfecte voorspelling), (2) een conditie waarin het netwerk willekeurige informatie kreeg over de volgende target (niet-informatieve voorspelling), en (3) een conditie waarin geen informatie werd verstrekt aan het netwerk (geen voorspelling).

De beste prestatie, gemeten door snelheid en nauwkeurigheid, werd geleverd door de netwerken die perfecte informatie over de volgende target kregen. Deze netwerken stuurden de robotarm naar de volgende target, nog voordat deze target actief werd. Zij leerden dus gebruik te maken van de informatie die aan ze werd verstrekt. De netwerken met niet-informatieve en geen voorspellingen scoorden minder goed. Zij evolueerden een strategie analoog aan die van mensen zonder expliciete kennis van de reeks: ze bewogen de robotarm naar het midden van de stimuli, op gelijke afstand van alle potentiële targets. Ook was zichtbaar dat de netwerken met niet-informatieve voorspellingen langzamer evolueerden dan de netwerken zonder voorspellingen. Het kost blijkbaar tijd om de willekeurige invoer te negeren. Vervolgonderzoek zou kunnen uitwijzen of dit vergelijkbaar is met proefpersonen die *denken* expliciete kennis over de reeks te bezitten, maar dit in feite niet hebben.