Publicatie : AG Connect
Datum : 30 sep 2021 Regio : Nederland Pagina : 41 Bereik : -1
cm2 : 1350 Frequentie : 10x per jaar AVE : € 12250,00 Oplage : 7.700

Net als software en websites moet ook kunstmatige intelligentie (AI) uitvoerig getest worden. Maar bij software is de verwachte output constant en goed te specificeren. AI-systemen zijn dynamisch en hebben niet eens altijd een 'ground truth'. Hoe test je dan zo'n systeem?

door Eveline Meijer beeld Shutterstock

AI TESTEN LIJKT OP EEN RIJEXAMEN:

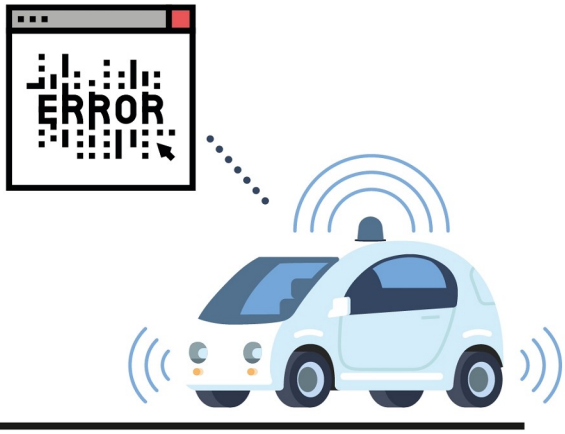
HET DRAAIT OM VERTROUWEN

HET TESTEN VAN SOFTWARE OF EEN WEBSITE IS BEHOORLIJK BINAIR: ontwikkelaars en testers geven aan wat de verwachte output bij bepaalde input is en controleren of de daadwerkelijke output overeenkomt. Daar komt vervolgens uit of de software slaagt of faalt.

AI is anders. Het heeft te maken met diverse soorten input en allerlei interacties met ofwel mensen ofwel andere systemen. "Je kunt nog steeds wel een verwacht resultaat specificeren, maar het echte resultaat kan daarvan afwijken en dat hoeft niet per se fout te zijn", vertelt Tom van de Ven, principal consultant innovation bij Sogeti. "Je moet dus in bandbreedtes denken: iets is goed

Publicatie : AG Connect
Datum : 30 sep 2021 Regio : Nederland Pagina : 41 Bereik : -1
cm2 : 1350 Frequentie : 10x per jaar AVE : € 12250,00 Oplage : 7.700

Zelfs de mensen die het algoritme trainen kunnen waarschijnlijk niet precies zien waar de fout zit



binnen een bepaalde bandbreedte."
De output kan bovendien veranderen als een scenario nog een keer doorgelopen wordt en het is geen gegeven dat die output óók correct is. Er kan later dus alsnog een fail uitkomen. Daarnaast komt het voor dat een systeem helemaal geen 'ground truth' heeft, weet dr. Tim Baarslag van het Centrum Wiskunde & Informatica (CWI) en de Universiteit Utrecht. "Soms weten we niet precies wat de juiste output is, en kunnen we alleen achteraf bepalen wat wel en wat niet werkt."

RIJEXAMEN

Om AI dan toch goed te kunnen testen, zijn diverse methodes bedacht. Eén daarvan draait om simulaties: een chatbot of een zelfrijdende auto krijgt allerlei situaties voorgeschoteld om te zien of-ie zich gedraagt zoals zou moeten.

"Je kunt het goed vergelijken met een rijexamen. Wij als mens moeten eerst een test doen voor we een auto mogen besturen. We beoordelen dus of je in staat bent om die auto te besturen en de rit tot een goed einde te brengen", legt Van de Ven uit. Het nadeel is natuurlijk het nooit mogelijk is om álle scenario's te testen. Net zoals geldt bij een rijexamen: "We creëren met een rijexamen het vertrouwen dat de bestuurder het er in de toekomst goed vanaf gaat brengen, ook in situaties die nog niet getest zijn. Voor AI's moet je iets vergelijkbaars ontwikkelen: vooraf genoeg scenario's bedenken die maken dat je er na deze tests op kan vertrouwen dat de AI ook in de toekomst goede resultaten oplevert."
Daarmee red je het niet bij het doorlichten van AI's met een heel grote vrijheidsgraad. Of iets een goed antwoord is, hangt in deze systemen van heel veel factoren af. In dit soort gevallen is het onder meer mogelijk om een 'fixed' model te bouwen. "Dan laat je het model eerst leren en op basis daarvan bouwt hij dan een neuraal netwerk. Dat netwerk fixeer je. Dus dan heeft het systeem een vast algoritme waarbinnen het acteert."
In dat soort gevallen is het niet nodig om testsituaties te bedenken, maar gaat het veel meer om de wiskunde achter het model. "Welke wiskundige algoritmes zijn gebruikt? Kunnen we bewijzen dat het gebruikte algoritme tot de juiste conclusies kan komen? En is de dataset die we gebruiken voor de training van een kwaliteit die hoog genoeg is?
Dat kun je vinden door steekproeven te doen. Op basis daarvan zet je een wiskundig vertrouwen in de oplossing op", aldus Van de Ven.

WEDSTRIJD TUSSEN AI'S

Baarslag test zijn AI-modellen ook nog op een andere manier. Hij ontwikkelt systemen die in allerlei omstandigheden kunnen onderhandelen, bijvoorbeeld over het kopen van een huis of over een

Publicatie : AG Connect
Datum : 30 sep 2021 Regio : Nederland Pagina : 41 Bereik : -1
cm2 : 1350 Frequentie : 10x per jaar AVE : € 12250,00 Oplage : 7.700

nieuwe baan. De algoritmes hebben als doel om de beste deal te leveren voor de gebruiker in kwestie. Maar wat die beste deal dan is en wat de beste manier is om die deal te krijgen, is niet precies duidelijk en kan per persoon verschillen.
De één wil meer salaris, de ander juist meer ouderschapsverlof. De ene onderhandelingspartner is wel gevoelig voor vleierij en de ander niet.
Om toch te achterhalen wat wel en niet werkt, wordt een internationale competitie opgezet: de Automated Negotiating Agent Competition (ANAC). Daar strijden onderhandelingsalgoritmes tegen elkaar, om te zien wie de beste is van dat jaar. De algoritmes kunnen in allerlei omgevingen en omstandigheden onderhandelen, wat betekent dat ze heel generiek moeten zijn. "Anders zou je bij elke onderhandeling een andere AI-oplossing nodig hebben en kun je ze niet vergelijken", verklaart Baarslag. Op basis van welke voorkeuren een algoritme in een bepaalde situatie de beste deal moet sluiten, wordt dus pas tijdens de competitie duidelijk.
De algoritmes moeten bovendien in verschillende rollen kunnen onderhandelen. "We laten ze alle zijdes van alle mogelijke onderhandelingen spelen, zonder dat ze kunnen onthouden wat ze hebben gespeeld. Ze worden dus iedere keer gereset. We kijken wat voor deals ze halen, waarbij ze ook moeten spelen tegen de andere algoritmes die er zijn. Zo krijg je een eerlijke ranking."

ZELFS DE DATASET IS EEN BLACK BOX

Mocht een systeem niet werken zoals bedoeld, dan moet het probleem worden opgelost. Bij software testing wordt de oorzaak in de code gezocht. Maar bij veel AI-systemen kan dat niet. "Zelfs de mensen die het algoritme trainen kunnen waarschijnlijk niet precies zien waar de fout zit. Dus dat moet je terug kunnen relateren aan de data die je gebruikt hebt om te leren. Dat is heel lastig", zegt Van de Ven.
"Als je bijvoorbeeld een AI-systeem hebt dat plaatjes herkent en dat gaat een keer mis, dan moet je terug naar de database met plaatjes die je gebruikt hebt voor de training om te zien wat er mis is gegaan. Misschien heeft het systeem een schoen in een bepaalde positie wel aangezien als een open jurk. Dat uitzoeken is heel tijdintensief. De black box gaat dus verder dan het algoritme en het stukje neurale netwerk dat je hebt. Voor de tester gaat dat zo ver dat zelfs de testdataset onderdeel is van de black box."
Toch is het vaak wel mogelijk om een root cause-analyse te doen. "Dat doe je door bijvoorbeeld plaatjes te zoeken die lijken op degene waar het mis ging, om te kijken of het daarbij ook misgaat. Gaat het bijvoorbeeld fout bij een plaatje van een bruin brood, dan kun je onderzoeken of dat bij alle broden zo is, of dat het specifiek om de kleur of vorm gaat."
Daarnaast kun je met het testen uitzoeken wat nog beter kan. Bij de competities waar Baarslag aan meedoet, worden algoritmes achteraf geanalyseerd.
Zo wordt onderzocht welke algoritmes het goed doen bij grote onderhandelingen met veel onderwerpen of juist bij de kortere onderhandelingen met mensen, en welke strategie goed werkt. "Maar testen is vaak maar een onderdeel van de cyclus. En als iets niet goed werkt, ben je vaak juist blij, want dat is een verbeterpunt waardoor je weer iets nieuws kunt maken en dat weer kunt testen.
De competitie is dus een soort test in een cyclus die ieder jaar herhaald wordt, waardoor de cyclus steeds beter wordt."

ALTIJD BLIJVEN TESTEN

Een groot nadeel van AI is dat één keer testen niet genoeg is, zelfs niet als dat heel uitvoerig gebeurt. Zeker zelflerende AI-systemen zijn dynamisch en kunnen uiteindelijk andere antwoorden gaan geven bij dezelfde input, die niet per se goed zijn. Daarom is het volgens Van de Ven belangrijk om constant te blijven testen. "Bij een AI-product of een AI-oplossing in een product moet je niet alleen vertrouwen krijgen dat het goed werkt, daarna moet je dat vertrouwen ook houden. Je wil dus constant een test laten lopen, die scenario's blijft afvuren op bijvoorbeeld een chatbot, om te zien of het systeem binnen de bandbreedte blijft antwoorden. Je moet constant een vinger aan de pols houden."

Jaarlijkse game AI-onderhandelen houdt algoritmes scherp