Generative Transformer technologie

Sako Arts is CTO bij de tech startup FruitPunch AI. FruitPunch houd zich bezig met AI for Good Challenges voor bedrijven en organisaties met behulp van hun community. Daarnaast ontwikkelt het bedrijf Challenge Based Education om deze community, maar ook bedrijven, te voeden met de state-of-the-art AI kennis en kunde. Hiervoor werkte Sako tweeëneenhalf jaar als AI Specialist bij softwarebedrijf Wolfpack IT. Sako heeft een achtergrond in informatica en hij is de eerste generatie van de Data Science in Engineering Master van de TU/e.

De technologie Natural Language Processing is gericht op het begrijpen van de menselijke taal, op een manier waarop mensen communiceren en hoe zij daarbij op elkaar reageren. OpenAI is een van de organisaties die veel onderzoek doet naar dit onderwerp en recent is een nieuwe technologie ontwikkeld: GPT-3 (Generative Pre trained Transformer 3). Deze techniek wordt steeds vaker toegepast en maakt gebruik van vooraf getrainde algoritmes om teksten te genereren en heeft in de afgelopen maanden veel belangstelling gekregen. Wat kan je met GPT-3 en waarom is de belangstelling in GPT-3 toegenomen?
"GPT-3 is een taalgeneratiemodel. Dat betekent dat het zelf zinnen, teksten of zelfs hele verhalen kan genereren. Je kunt het model bijvoorbeeld een startzin of alinea geven en het zal dan zelf een verhaal schrijven dat een vervolg van de gegeven input is. GTP-3 is de derde generatie van de OpenAI modellen en is bijna akelig goed in deze taak. De teksten zijn niet tot nauwelijks te onderscheiden van teksten die door echte mensen zijn geschreven. Wat dit indrukwekkend maakt is niet alleen dat dit model zo'n goed begrip heeft van de taal dat het een vrijwel perfecte grammatica en een uitgebreide vocabulaire heeft, maar vooral dat het een coherent en logisch verhaal weet te genereren.

Hierin verwerkt het zowel feiten als meningen die consistent blijven door het stuk heen. Dus het model is zowel op de hoogte van een onnoemelijk aantal feiten over onze wereld als dat het begrip lijkt te hebben van de verschillende perspectieven die wij mensen kunnen hebben op een onderwerp. Het woordje 'lijkt' in die laatste zin is overigens wel essentieel... Feitelijk heeft het model gewoon door zo veel verschillende teksten heen gekeken in het trainingsproces dat het voor vrijwel elk perspectief een context heeft gecreëerd, en het weet die verschillende contexten ook aan elkaar te relateren door tekstuele overeenkomsten.

HIERDOOR KAN HET DUS VERSCHILLENDE PERSPECTIEVEN RECONSTRUEREN

Het model heeft dan ook verschillende maanden dag in dag uit getraind terwijl het een heel datacenter aan computerkracht gebruikt. Het resulterende model is dan ook letterlijk een miljoen keer groter dan andere veelgebruikte modellen in de industrie. Hierdoor kan het dus verschillende perspectieven reconstrueren en zelfs verrijken met andere perspectieven. Het kan echter niet een heel nieuw eigen perspectief vormen. Daar zit een belangrijk verschil met mensen."

GPT-3, wordt als één van de meest interessante en belangrijke AI-systemen gezien. OpenAI opereert vanuit een nonprofit gedachte en dit maakt dat de technologie toegankelijk is voor een groot publiek. Recent heeft Microsoft als partner en investeerder een belang gekregen in de verdere ontwikkeling van GPT-3. Wat is je mening omtrent deze ontwikkeling?
"Zelf geloof ik niet zo in 'evil corporations that want to rule the world'. Een groot bedrijf dat een aandeel koopt en een relatie aangaat met een veelbelovende technologische partner is in mijn ogen dagelijkse gang van zaken en hier heb ik geen specifieke mening over.

Ik vind het daarentegen juist een hele logische move. Zoals eerder gesteld kost het trainen van een model als GTP-3 heel veel computerkracht. Als je standaard Cloud Computing capaciteit zou inkopen om dit model te trainen ben je naar schatting 4,6 miljoen dollar kwijt. OpenAI bezit en onderhoudt daar nu zelf hardware voor, wat natuurlijk niet hun kerntaak is - dat is het ontwikkelen van AI. Microsoft Azure heeft hier wel een focus liggen en een samenwerking zoals ze nu aangaan zal ervoor zorgen dat dit soort onderzoek voor OpenAI betaalbaar blijft."

Als we kijken naar de toepasbaarheid van GTP-technologie, beperkt de technologie zich dan tot het trainen van algoritmes om teksten te generen, of is de technologie ook toepasbaar bij genereren van pixels om afbeeldingen te voorspellen?
"Het specifieke GTP-3 model is getraind op teksten en zal niets anders kunnen dan dat. Dit geldt echter niet voor de onderliggende Generative Transformer technologie en het trainen ervan op een dergelijk grote schaal. Deze kun je voor veel meer toepassingen inzetten, bijvoorbeeld pixels. Sterker nog, OpenAI heeft dit onlangs gedaan met hun DALL-E model, dat vergelijkbare technologie gebruikt om compleet nieuwe afbeeldingen te genereren op basis van een tekstuele beschrijving."

Aanvullend op de vorige vraag: zie je wellicht mogelijkheden dat GTP- technologie in de toekomst wellicht AI programmacode kan schrijven?
"Er zijn al verschillende onderzoeken bezig om code te genereren, er is zelfs al software die je als programmeur kan gebruiken en die code automatisch aanvult zoals je telefoon dat ook doet voor je appjes. Few-shot learning technologie en transformers zijn hier erg voor geschikt.

Echter, zoals eerder gezegd, deze modellen kunnen contexten recreëren en combineren, maar niet zelf denken. Volledig nieuwe AI-programma's schrijven zal deze technologie dan ook nooit kunnen. Mocht dat überhaupt mogelijk zijn zullen we daar nog een totaal nieuwe doorbraak voor nodig hebben. Daar bovenop is menselijke taal significant makkelijker voor een dergelijk model dan programmacode. Waar je bij taal heel veel verschillende dingen kan genereren die allemaal niet fout zijn, komt dat een stuk nauwer bij programmacode. Je hoeft maar een kleine fout in de code te maken en het hele programma crasht."

Veel mensen nemen informatie op internet voor waar aan. Hoe zie jij het gevaar van het gebruik van GPT ten aanzien van fake news/trollen en dergelijke?
"Ik zie het inmiddels niet meer als een gevaar maar als een gegeven, die streep zijn we allang voorbij. Dit soort technologie maakt het voor een kwaadwillende wel veel makkelijker om op grote schaal mis-informatie, die lastig te onderscheiden is van het echte, te produceren en te verspreiden. Dit konden ze echter ook al door een heel leger aan schrijvers aan te nemen.

WE ZULLEN ALS MENSEN GEWOON KRITISCHER MOETEN WORDEN OP WAT WE LEZEN

Het feit dat ze het nu goedkoper en sneller kunnen verandert de zaak niet echt naar mijn mening. We zullen als mensen gewoon kritischer moeten worden op wat we lezen en waar het vandaan komt. Vanuit een zeer optimistisch oogpunt zorgt dit er juist voor dat mensen kritischer worden omdat ze weten dat het ook gegenereerd kan zijn. Ik weet niet of ik het zelf geloof, maar ik kan hopen..."

Als we kijken naar de financiële dienstverlening, kun je dan op basis van de nieuwe GPT-3 technologie aangeven of deze technologie disruptief zal zijn en zo ja, kun je dan een voorbeeld van een toepassing geven?
"Voor taalgeneratiemodellen zie ik geen specifieke toepassingen voor de financiële dienstverlening. Natuurlijk zijn er voordelen uit te halen maar die zijn relevant voor veel industrieën en niet uniek voor de financiële. Überhaupt is generatieve technologie niet zeer van toe-passing in financiën. Daar moet je het meer hebben van voorspellingen en modellering van de werkelijkheid. Nu kun je een generatief model ook zien als een modelering van de werkelijkheid, maar deze modellen zijn specifiek goed in het garneren van meerdere mogelijke werkelijk-heden, niet zo zeer in een specifieke, namelijk onze toekomstige werkelijkheid."

Als teksten met AI worden gemaakt, hoe betrouwbaar zijn die dan en welke risico's loop je als bedrijf?
"Betrouwbaar zijn ze op geen enkele manier, je kunt namelijk nauwelijks controleren wat er uit het model komt. Je kunt er bij GTP-3 inmiddels aardig op vertrouwen dat het een grammaticaal consistente zin construeert en dat het de woorden op een juiste manier gebruikt. Maar zelfs dit is geen gegeven. Wat het model genereert en of dit waar is heb je dus geen controle over.

Sterker nog, een dergelijk model kan ook prima racistische of anders offensieve teksten genereren, dit is erg afhankelijk van de teksten waarmee het model is getraind. Het risico wanneer je zo'n model inzet binnen je bedrijf is dus dat het misinformatie verspreidt of gebruikers beledigt. Neem een voorbeeld aan Microsofts Twitter bot (https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist) van een paar jaar geleden." ■