Publicatie : PC-Active
Datum : 31 mrt 2021 Regio : Nederland Pagina : 90 Bereik : -1
cm2 : 2128 Frequentie : 6x per jaar AVE : € 28214,00 Oplage : 65.000

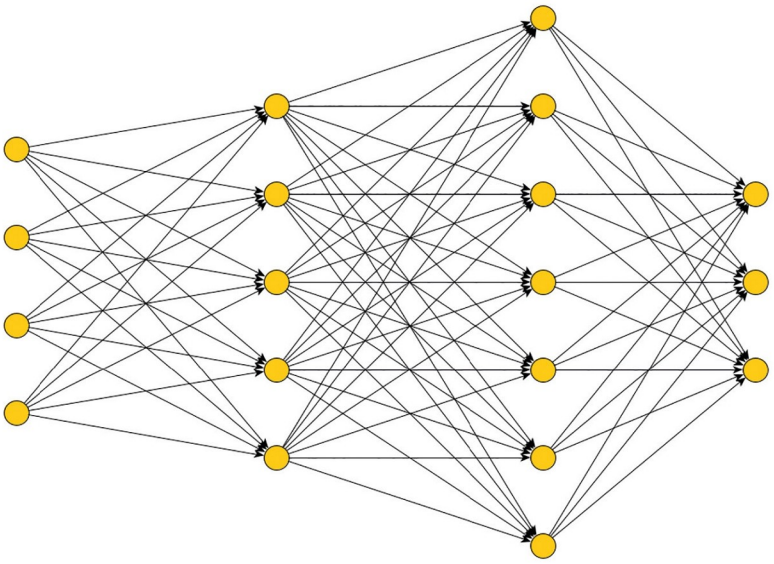
KOEN VERVLOESEM

DENKWERK

ENERGIEZUINIGERE NEURALE NETWERKEN DOOR PULSEN

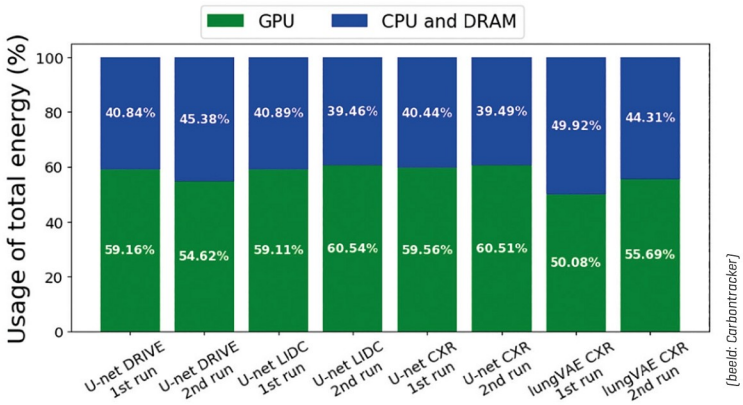
Kunstmatige intelligentie komt de laatste tijd vaak negatief in het nieuws. Het trainen van neurale netwerken voor *deep learning* is immers energieverslindend. Nederlandse onderzoekers zijn nu tot een doorbraak gekomen die toepassingen zoals spraakherkenning een factor 100 tot 1000 energiezuiniger moet maken.

Als we onze hersenen als technologie zouden beschouwen, zou elke ingenieur onder de indruk zijn: met een verbruik van amper 20 watt slaagt ons brein erin om talloze gevarieerde en complexe taken uit te voeren, zoals spraak en beeld herkennen, navigeren in omgevingen waar we nog nooit geweest zijn, nieuwe vaardigheden leren en redeneren over abstracte zaken. Het is dan ook geen wonder dat onze hersenen al sinds jaar en dag als inspiratie dienen om computers 'intelligentie' te geven. Een belangrijke aanpak in machinaal leren vormen (kunstmatige)



Een volledig verbonden neuraal netwerk met twee verborgen lagen

Publicatie : PC-Active
Datum : 31 mrt 2021 Regio : Nederland Pagina : 90 Bereik : -1
cm2 : 2128 Frequentie : 6x per jaar AVE : € 28214,00 Oplage : 65.000



(beeld: Carbontracker)

De gpu is goed voor meer dan de helft van het energieverbruik van het trainen van een neuraal netwerk

neurale netwerken. Ze bootsen de werking van de hersenen na, die een biologisch neuraal netwerk vormen: een kluwen van talloze verbindingen tussen neuronen (hersencellen).

NEURALE NETWERKEN

Een kunstmatig neuraal netwerk bestaat meestal uit meerdere lagen: een invoerlaag van neuronen die de invoer van een probleem voorstellen (bijvoorbeeld de pixels in een foto), een uitvoerlaag van neuronen die de oplossing van het probleem voorstellen (bijvoorbeeld: het is een foto van een hond) en één of meer tussenliggende (verborgen) lagen die berekeningen uitvoeren - die bijvoorbeeld vacht, grootte, aantal poten enzovoorts herkennen.

Een neuraal netwerk programmeer je niet door expliciet aan te geven hoe het een probleem moet oplossen; je 'traint' het door het vele voorbeelden van een probleem te geven. De parameters van alle neuronen van het neurale netwerk convergeren door die training dan naar de juiste waarden, zodat het de taak leert uit te voeren.

Vooral *deep learning* maakt het laatste decennium furore in de wereld van *machine learning*. Bij deep learning maak je gebruik van een neuraal netwerk met een groot aantal lagen tussen invoer en uitvoer. Door dit grote aantal lagen zijn eindelijk heel complexe taken mogelijk.

Een neuraal netwerk als GPT-3 (zie Denkwerk in het vorige nummer van PC-Active) gebruikt zo'n honderd lagen. Als je het volledig zelf zou willen trainen, kijk je aan tegen enkele miljoenen euro's aan kosten om gpu-rekenkracht in de cloud te huren.

GROOT ENERGIEVERBRUIK

Niet alleen de kostprijs van een netwerk voor deep learning is hoog, ook het energieverbruik. Twee Deense studenten hebben een tool ontwikkeld om de CO_2-voetafdruk van het trainen van een model voor deep learning te schatten: carbontracker (https://github.com/lfwa/carbontracker). Zij schatten dat het energieverbruik om GPT-3 te trainen even hoog is als het jaarlijkse energieverbruik van 126 Deense huizen en evenveel CO_2 uitstoot als 700.000 km autorijden. Terwijl AI enkele jaren geleden nog werd geprezen omdat we er slimme oplossingen mee zouden vinden voor de klimaatopwarming, zou het op deze manier juist sterk gaan bijdragen aan de klimaatopwarming.

Ook het uitvoeren van het neuraal netwerk zodra het is getraind, kost nog altijd veel meer energie dan de menselijke hersenen. Als we een neuraal netwerk ter grootte van de hersenen zouden maken, zou dat enkele megawatt verbruiken, het equivalent van de energie die een kleine biomassacentrale of waterkrachtcentrale produceert. Onze kunstmatige neurale netwerken zijn een factor miljoen minder efficiënt dan hun biologische evenknieën.

PULSEN IN HET BREIN

De grote uitdaging is dus om die genoemde factor miljoen te verlagen. Een van de onderzoekers die zich daarop heeft gestort, is prof. dr. Sander M. Bohté, die in 2016 mee de onderzoeksgroep Machine Learning van het Centrum Wiskunde & Informatica (CWI) in Amsterdam heeft opgericht. Volgens hem zijn de klassieke kunstmatige neurale netwerken niet voldoende geïnspireerd door hun biologische evenknieën.

Neuronen in onze hersenen communiceren immers met pulsen. Gemiddeld sturen ze één puls per seconde, maar neuronen zijn niet continu ▶

Publicatie : PC-Active
Datum : 31 mrt 2021 Regio : Nederland Pagina : 90 Bereik : -1
cm2 : 2128 Frequentie : 6x per jaar AVE : € 28214,00 Oplage : 65.000

[foto: Inge Hoogland]

Professor Sander Bohté van het CWI onderzoekt al meer dan 20 jaar gepulste neurale netwerken

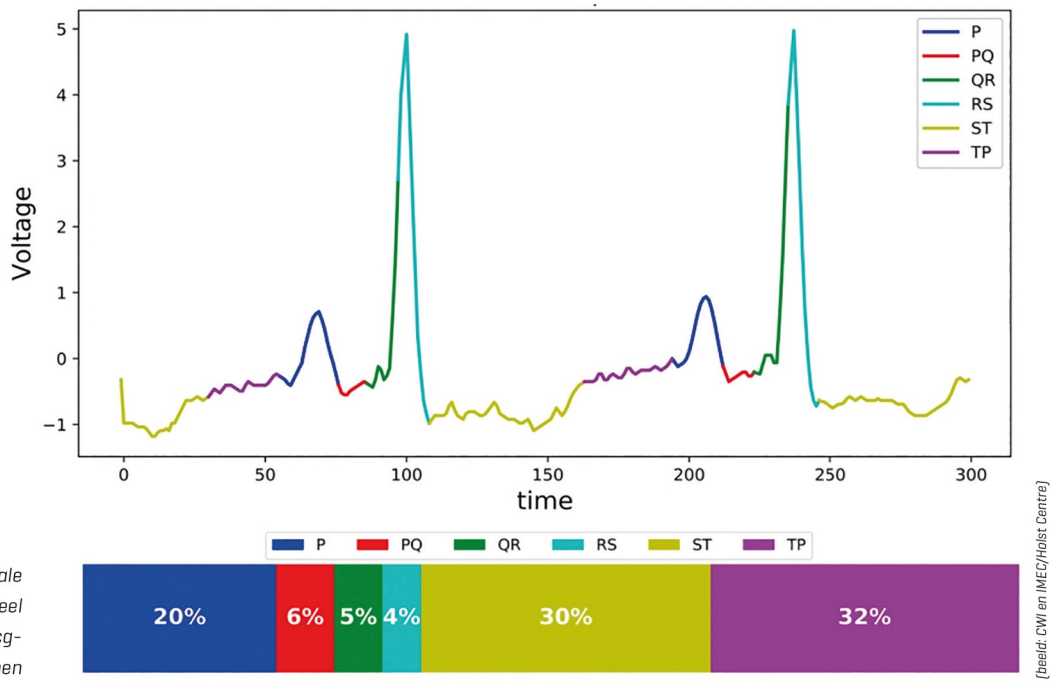
actief. Soms doen ze een seconde niets en soms vuren ze tien keer in een seconde. Op het moment dat ze geen puls sturen, verbruiken ze ook geen energie. Dat aspect van die pulsen hebben de klassieke neurale netwerken altijd genegeerd. Continue wiskundige signalen zijn nu eenmaal wiskundig makkelijker te hanteren dan de discontinue pulsen van onze hersenen.

GEPULST NEURAAL NETWERK

Toch is er al de hele geschiedenis van AI ook een aanpak geweest om neurale netwerken te modelleren met pulsen. Dat noemen we een gepulst neuraal netwerk (*spiking neural network*). Maar deze netwerken, die dus net zoals biologische neuronen geen energie verbruiken wanneer er niets gebeurt, zijn door hun wiskundige complexiteit nooit echt doorgebroken.
Al sinds zijn doctoraat van rond de eeuwwisseling doet Bohté onderzoek naar gepulste neurale netwerken, om te proberen neurale netwerken energie-efficiënter te maken. Onlangs ontwikkelde hij samen met onderzoekers Bojian Yin van het CWI en Federico Corradi van het Eindhovense onderzoekscentrum IMEC/Holst Centre een nieuw algoritme voor gepulste neurale netwerken.

DUBBELE DOORBRAAK

Het algoritme bevat twee doorbraken. Niet alleen hoeven de neuronen in het netwerk veel minder met elkaar te communiceren, daarnaast hoeft elk neuron ook nog eens minder te rekenen. In combinatie zorgen beide doorbraken ervoor dat de gepulste neurale netwerken volgens theoretische berekeningen een factor honderd energiezuiniger zijn dan de beste hedendaagse klassieke neurale netwerken.
Voorlopig is het algoritme maar geschikt voor gepulste neurale netwerken tot zo'n duizend neuronen. De klassieke aanpak om neurale netwerken te trainen werkt immers niet voor gepulste neurale netwerken. Maar toepassingen zoals spraakherkenning, de classificatie van elektrocardiogrammen (ecg) en het herkennen van gebaren zijn zelfs met deze beperking al mogelijk. Toch zal het nog een fikse uitdaging worden om de netwerken op te schalen naar 100.000 of een miljoen neuronen, wat voor complexe toepassingen nodig is.



Gepulste neurale netwerken kunnen heel energie-efficiënt ecg-golfvormen herkennen

[beeld: CWI en IMEC/Holst Centre]

Publicatie : PC-Active
Datum : 31 mrt 2021 Regio : Nederland Pagina : 90 Bereik : -1
cm2 : 2128 Frequentie : 6x per jaar AVE : € 28214,00 Oplage : 65.000

In hun paper *Effective and Efficient Computation with Multiple-timescale Spiking Recurrent Neural Networks* (https://arxiv.org/abs/2005.11633) beschrijven de onderzoekers hun aanpak en de resultaten die ze behaalden op enkele typen problemen. Afhankelijk van de taak en het type klassiek neuraal netwerk waarmee wordt vergeleken, zijn de gepulste neurale netwerken 28 tot 1900 keer energiezuiniger.

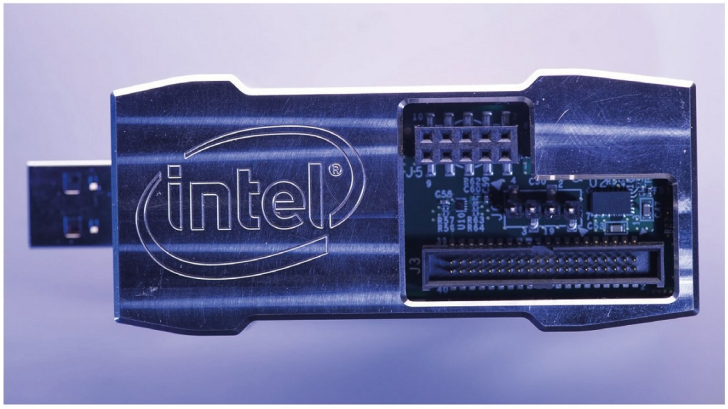
GESPECIALISEERDE CHIPS

Net als voor klassieke neurale netwerken is er voor gepulste neurale netwerken al hardware ontwikkeld die deze efficiënt uitvoert, want op normale processoren draaien de algoritmes veel te traag. Vaak heeft men het dan over *neuromorfische* of *cognitieve* computers, omdat ze nog dichter de werking van de menselijke hersenen benaderen dan versnellerhardware voor klassieke neurale netwerken. Bekend is SpiNNaker (http://apt.cs.manchester.ac.uk/projects/SpiNNaker/), waarvan de naam staat voor Spiking Neural Network Architecture. Het is een supercomputerarchitectuur van de universiteit van Manchester waarbij elke chip 16.000 menselijke neuronen *real-time* kan simuleren. SpiNNaker wordt gebruikt in het Human Brain Project (https://www.humanbrainproject.eu) om met een miljoen van deze processorkernen meer inzicht te krijgen in de werking van onze hersenen.

NEURAAL NETWERK IN EEN USB-STICK

Ook IBM ontwikkelt chips met gepulste neurale netwerken. In 2014 produceerde Big Blue de TrueNorth-chip, die een miljoen neuronen kan simuleren met een verbruik van slechts 70 mW. En in 2017 bracht Intel de chip Loihi uit, die simuleert 130.000 neuronen en zou volgens Intel duizend keer zo snel en tienduizend keer zo energie-efficiënt zijn als cpu's.

Intel heeft diverse systemen op basis van de Loihi-chip ontwikkeld, voornamelijk voor onderzoekers. Dat varieert van de usb-stick Kapoho Bay met een of twee Loihi-chips tot het grootschalige systeem Pohoiki Springs dat met 768 van deze chips meer dan honderd miljoen neuronen kan simuleren. Het is de bedoeling dat deze energiezuinige chips uiteindelijk in apparatuur terechtkomen die bijvoorbeeld een hartslag van een persoon real-time op onregelmatigheden kan analyseren of ziektes kan herkennen op basis van de adem van een persoon.



[foto: Intel]

De Loihi-chip van Intel past zelfs in een usb-stick die 260.000 neuronen simuleert

MEER LOKALE INTELLIGENTIE

Uiteindelijk zullen energiezuinigere neurale netwerken tot een verschuiving leiden in de plaats waar de berekeningen gebeuren. Omdat het rekenwerk van zogenoemde 'slimme' toepassingen momenteel nog zo zwaar en energieverslindend is, gebeurt het vaak in grote servers in de cloud. Denk bijvoorbeeld aan spraakherkenning: je zegt iets tegen je Google Home, de audio-opname gaat naar een server van Google, wordt daar geanalyseerd, Google beslist wat het je gaat antwoorden, stuurt de audio naar je Google Home en die spreekt het antwoord uit.

Dat brengt allerlei problemen met zich mee. Zo introduceert het naar de cloud sturen van de data en weer ontvangen van het resultaat altijd een vertraging. En als de netwerkverbinding uitvalt, verliest je apparaat al zijn intelligentie. Een ander probleem met deze aanpak is privacy: je stuurt allerlei persoonsgegevens (wat je in huis zegt) naar Google.

Maar als de neurale netwerken dankzij nieuwe algoritmes en nieuwe hardware een factor honderd of meer energiezuiniger worden, kunnen heel wat toepassingen die momenteel nog de cloud nodig hebben, gewoon op een smartphone of smartwatch draaien. En zo kunnen gepulste neurale netwerken wel eens tot robuustere en meer privacy-vriendelijke AI leiden. ■