Artificiële neurale netwerken als psychiatrisch instrument

L. Mertens, J. Vennekens, H. Op de Beeck, E. Yargholi, J. Van den Stock

Achtergrond Artificiële intelligentie (AI) evolueerde enorm het voorbije decennium en kent steeds meer toepassingsgebieden, ook in de psychiatrie. AI omvat verschillende modaliteiten, waaronder artificiële neurale netwerken (ANN's), verwijzend naar computermodellen die gebaseerd zijn op de werking van het brein. Hoewel ANN's sinds de jaren 50 beschreven zijn, werden ze pas 'mainstream' sinds een tiental jaren. Het feit dat ANN's inspiratie halen uit de werking van het brein doet de vraag rijzen of ze gebruikt kunnen worden om het (dis)functioneren van het brein te simuleren. Deze vraag heeft geleid tot de opkomst van het domein 'computationele psychiatrie'.

Doel Een toegankelijke introductie bieden tot artificiële neurale netwerken, en de toepassingsmogelijkheden in de hedendaagse psychiatrische praktijk.

Methode Literatuurbespreking met voorbeelden.

Resultaten Met enkele concrete voorbeelden schetsen we wat artificiële neurale netwerken zijn en hoe ze gebruikt kunnen worden om mechanismen in de hersenen te modelleren. We bespreken achtereenvolgens ANN's als model van het menselijk visueel systeem, als model van prosopagnosie en als model van auditieve hallucinaties en ten slotte als model van autismespectrumstoornis. Ook beschrijven we een aantal beperkingen van deze aanpak.

Conclusie Een computermodel dat het volledige brein nabootst, is momenteel nog een uitdaging, maar huidige modellen kunnen helpen met het testen van hypotheses over mechanismen die aan de basis liggen van neuropsychiatrische stoornissen.

In het relatief jonge onderzoeksveld computationele psychiatrie gebruikt men computermodellen om hersenprocessen te modelleren, en zo voorbij de symptomen (extern waarneembaar gedrag), een beter inzicht te verkrijgen in de (interne, onzichtbare) biologische oorzaak van neuropsychiatrische verschijnselen. Doel hierbij is hypotheses te testen over hoe een bepaald neurologisch deficit kan leiden tot een bepaald symptoom, zonder rechtstreeks te moeten interveniëren in de hersenen.

Artificiële neurale netwerken

Een specifiek type computermodel, geïnspireerd door de werking van de hersenen en aangewend voor dit doel, is het *artificiële neurale netwerk* (ANN). De bouwsteen van een ANN is het *artificiële neuron*: een wiskundige constructie die verschillende getallen als invoer neemt en hiermee een wiskundige berekening uitvoert om zo een uitvoerwaarde, of *activatie*, te verkrijgen die op zijn beurt doorgegeven wordt als invoer aan een volgend neuron. Met de verbinding tussen twee neuronen wordt een *gewicht* geassocieerd, dat bepaalt hoeveel activatie van het ene neuron effectief doorstroomt naar het andere, en dat zowel positief (exciterend) als negatief (remmend) kan zijn. In wezen is een ANN een gigantische wiskundige berekening.

Een eenvoudig type netwerk is het meerlagige perceptron, waarbij neuronen in drie of meer opeenvolgende lagen worden gegroepeerd: een inputlaag, een outputlaag, en daartussen zogenaamde 'verborgen' lagen (zie figuur 1). Elk neuron uit een laag is verbonden met elk neuron uit de volgende laag; er zijn geen verbindingen tussen neuronen uit eenzelfde laag. Informatie stroomt dus laag per laag van input naar output. De manier waarop neuronen gecombineerd worden (hoeveel lagen, grootte van de lagen, etc.), noemt men de *architectuur* van het model.

AUTEURS

Laurent Mertens, promovendus Computerwetenschappen, KU Leuven, dept. Computerwetenschappen, Leuven.
Joost Vennekens, hoofddocent Computerwetenschappen, KU Leuven, dept. Computerwetenschappen, en Flanders Make, KU Leuven.
Hans Op de Beeck, gewoon hoogleraar Psychologie, KU Leuven, Onderzoekseenheid Brein & Cognitie, Leuven.
Elahe' Yargholi, postdoc Biomedische technologie, KU Leuven, Onderzoekseenheid Brein & Cognitie, Leuven.
Jan Van den Stock, hoofddocent Neuropsychiatrie, UZ Leuven, Leuven Brain Institute, Leuven.

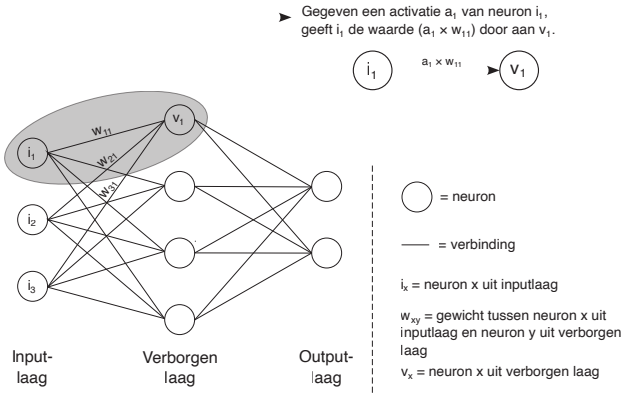
Correspondentie
Laurent Mertens (laurent.mertens@kuleuven.be).

Geen strijdige belangen meegedeeld.

Het artikel werd voor publicatie geaccepteerd op 18-10-2023.

Citeren
Tijdschr Psychiatr. 2023;65(10):646-650

Figuur 1. Model bestaande uit meerdere lagen neuronen, waarbij elk neuron verbonden is met alle neuronen uit de vorige laag, maar neuronen uit eenzelfde laag niet met elkaar verbonden zijn



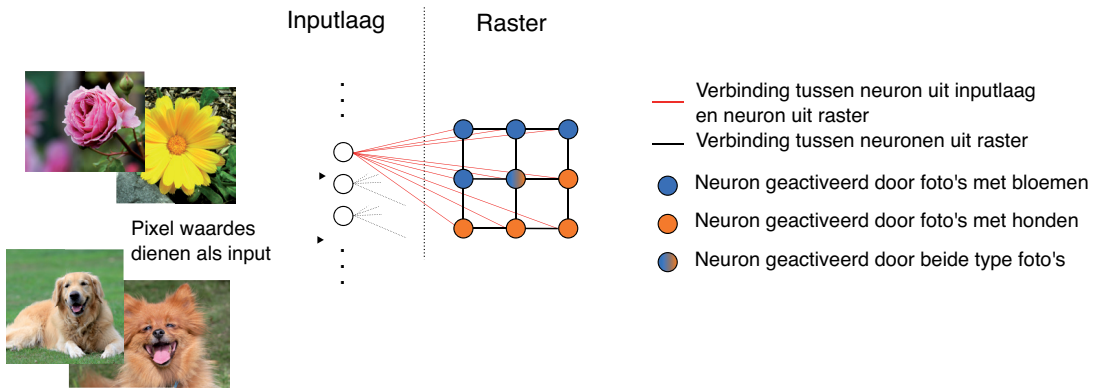
Toepassingen van ANN's

ANN's kennen vele toepassingen die hoofdzakelijk in twee categorieën onderverdeeld kunnen worden: *regressie* en *classificatie*.
Bij regressie heeft het netwerk doorgaans slechts één outputneuron en dient het een getalwaarde te voorspellen. Zo kan de score die een persoon morgen zal halen op een stemmingsvragenlijst voorspeld worden, gegeven als netwerkinput de huidige score, een kwantificatie van actuele psychosociale stressoren en sterkte van sociaal netwerk. De door het netwerk voorspelde score wordt dan uitgelezen via het outputneuron.
Bij classificatie probeert men met elke input een correct label te associëren uit een vaste set labels. Het netwerk heeft evenveel outputneuronen als er labels zijn; elk outputneuron vertegenwoordigt één label, en de numerieke waarde van een specifiek outputneuron wordt geïnterpreteerd als de probabiliteit, volgens het netwerk, dat het corresponderende label het juiste is voor de input in kwestie. Stel, we hebben een verzameling hersenscans van patiënten met een neurocognitieve stoornis (NCS) of depressie op latere leeftijd (DLL). Er zou een ANN geconstrueerd kunnen worden waarbij de input bestaat uit de intensiteitswaarden van de voxels, en twee outputneuronen waarbij één het label 'NCS' en het ander 'DLL' representeert. Het doel is dan dat bij input van hersenscans het correcte outputneuron maximaal geactiveerd wordt.

Hoe leren ANN's?

ANN's kunnen niet autonoom leren, ze dienen expliciet getraind te worden. Tijdens het trainen worden de gewichten van het netwerk zodanig afgesteld dat het netwerk het gewenste resultaat berekent.
Een manier om dit te doen is via *supervised learning*, waarbij een model getraind wordt met een *trainingsset*. Als we terugkeren naar onze NCS/DLL-classifier, zou dit een verzameling hersenscans zijn waarvan het label bekend is. Tijdens de trainingsfase laat men het netwerk een voorspelling maken voor elke scan, en vergelijkt deze met het juiste antwoord. Deze informatie wordt op wiskundig onderbouwde wijze aangewend om de gewichten van het model aan te passen zodat de voorspelling (dat wil zeggen de activaties van de outputneuronen) dichter bij het gewenste antwoord komt te liggen. Wanneer goede resultaten behaald worden op deze trainingsset, kan men het model aanwenden om voorspellingen te maken voor scans van nieuwe patiënten bij wie de diagnose nog niet gesteld is.
Bij *unsupervised learning* daarentegen wordt een grote hoeveelheid data aan een model gegeven, en is het doorgaans de bedoeling dat dit erin slaagt om intern vergelijkbare datapunten te clusteren, wat zich uit in een vergelijkbare netwerkrespons. Bij een *self-organizing map* (SOM) bijvoorbeeld is een laag inputneuronen ver-

Figuur 2. Illustratie van een self-organizing map



Dit hypothetische model slaagt erin om foto's van bloemen te onderscheiden van foto's van honden; beide types foto's activeren een andere neuroncluster met slechts 1 gemeenschappelijk neuron.

bonden met een raster van outputneuronen waarbij elk outputneuron verbonden is met al zijn buren (zie figuur 2). Het model wordt getraind zodanig dat vergelijkbare inputs (bijv. foto's van ófwel bloemen, ófwel honden) grotendeels dezelfde cluster outputneuronen activeren, en de overlap tussen neuronclusters zo klein mogelijk is. In het navolgende lichten we enkele toepassingen toe van het gebruik van ANN's als hersenmodel.

ANN's als model van het menselijk visueel systeem

Convolutionele neurale netwerken (CNN's), geïnspireerd door het menselijk visueel systeem, zijn erg populair voor beeldanalyse, bijv. het classificeren van afbeeldingen, objectherkenning (detecteren en benoemen van objecten op een foto) of segmentatie (alle pixels die eenzelfde entiteit uitmaken gelijk inkleuren). CNN's gebruiken filters die over een inputafbeelding 'geschoven' worden (vergelijkbaar met receptieve velden van biologische neuronen in het visueel systeem), waarbij op elke plek gekeken wordt naar de overeenkomst tussen afbeelding en filter, om zo representatieve visuele features (bijv. een rand) te extraheren. Deze filters worden weer in sequentiële lagen georganiseerd (zie figuur 3). Zowel in een aantal populaire CNN's als in het menselijk visueel systeem lijken de vroege lagen gevoelig voor eenvoudige geometrische eigenschappen (lijnen en randen, etc.), en worden de eigenschappen stelselmatig complexer in de diepere lagen (ANN)/verder in de ventrale visuele stroom (ventrotemporale cortex).

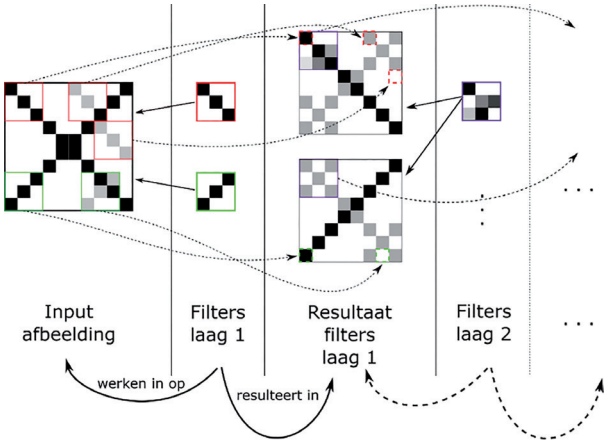
ANN-model van prosopagnosie

Prosopagnosie betreft een deficit in gezichtsherkenning, dat gelinkt is aan de *fusiform face area* (FFA) in de gyrus fusiformis. Om schade aan de FFA te modelleren werd een SOM getraind om afbeeldingen van 4 gezichten, een olifant en 5 objecten te encoderen in een raster van 20 x 20 neuronen. Vervolgens werd een classifier getraind die de SOM-representaties associeerde met de juiste labels ('gezicht 1', 'olifant', etc.). De SOM slaagde erin om de gezichten en overige klassen topologisch van elkaar te scheiden. Vervolgens werden de artificiële neuronen die de gezichten encodeerden, uitgeschakeld, waarna de classifier niet langer in staat was om deze correct te herkennen, terwijl herkenning van de overige klassen intact bleef.[1] Aldus levert dit model een hypothese over de biologische oorzaak van prosopagnosie.

ANN-model van hallucinaties

Een pathogenetische hypothese van auditieve hallucinaties bij schizofrenie stelt dat deze gelinkt zijn met excessieve synaptische *pruning* tijdens de adolescentie. Deze hypothese werd getest met een ANN bestaande uit 4 lagen: een inputlaag met 25 neuronen, een verborgen laag, een outputlaag met 43 neuronen en, cruciaal, een 'geheugen'-laag.[2] Voor een vocabularium van 30 termen waar éénvoudige zinnen mee gemaakt konden worden, werd met elke term een vector van 25 getallen geassocieerd die als netwerkinput diende, alsook een vector van 43 getallen die de gewenste netwerkoutput voorstelde. Het netwerk werd getraind met 256 zinnen die woord voor woord aan de inputlaag aangeboden werden, met als doel de inputvector om te zetten in de overeenstemmende outputvector. De verborgen laag kreeg als input een combinatie van de inputlaag en de geheugenlaag, die telkens een kopie bevatte van de toestand van de verborgen laag bij de verwerking van de vorige term. Hierbij slaagde het netwerk erin om informatie uit de geheugenlaag te gebruiken om de huidige term correct te identificeren.
Vervolgens werd het netwerk getest met 23 opeenvolgende zinnen, telkens gescheiden door 5 'nul'-tekens. Er werd gekeken naar hoeveel woorden het netwerk correct en foutief omzette, alsook hoe vaak het netwerk een woord voorstelde wanneer er geen input ('nul') was.

Figuur 3. Illustratie van het convolutie-mechanisme van een convolutioneel neuraal netwerk



De filters uit laag 1 werken in op de inputafbeelding. Op elke positie wordt de overeenstemming tussen filter en afbeelding gemeten, wat een nieuwe pixel oplevert. Hoe beter de overeenkomst, hoe donkerder de pixel. Elke filter genereert een nieuwe afbeelding, die door alle filters in de volgende laag afgetast wordt.

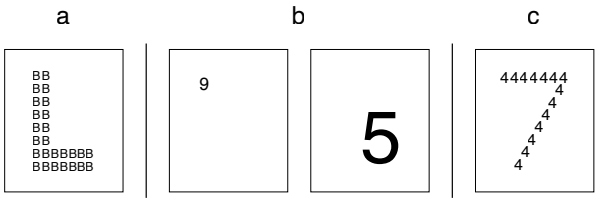
Dit laatste fenomeen werd beschouwd als een hallucinatie. Tevens werd geëxperimenteerd met het verbreken van netwerkverbindingen (als artificieel analoog van pruning). Het aantal correcte omzettingen verminderde afhankelijk van het aantal geprunede neuronen uit de inputlaag, maar er werden geen hallucinaties waargenomen. Wanneer echter de geheugenlaag gepruned werd, resulteerde dit eerst in verbeterde classificatie, maar bij excessief prunen begon het netwerk te 'hallucineren'. Dit ondersteunt de hypothetische associatie tussen synaptische pruning en hallucinaties.

ANN-model van autismespectrumstoornis (ASS)

ASS kan geassocieerd zijn met een excessieve focus op details, waarbij sneller lokale patronen dan globale herkend worden. Met een ANN werd de hypothese getest dat dit een gevolg zou kunnen zijn van een verstoord evenwicht tussen exciterende en inhiberende neuronen in de visuele cortex.[3] Hierbij werd getracht de resultaten van een gedragsexperiment na te bootsen waarbij participanten met en zonder ASS een beeld getoond werd van een grote letter (globaal patroon) die opgebouwd was uit een kleinere letter (lokaal patroon), zoals weergegeven in figuur 4a. Vervolgens werd hun gevraagd de grote dan wel kleine letter te benoemen. Personen met ASS waren sneller om de kleine letter te benoemen, tegenovergesteld aan de controlegroep.

Het ANN was specifiek ontwikkeld voor de extractie van visuele features, en bestaat uit een opeenvolging van S- (simple) en C- (complex) lagen. In de S-lagen worden eenvoudige patronen gedetecteerd, die door de C-lagen verwerkt worden in complexere features. Belangrijk is dat binnenin de S-lagen een remmend mechanisme

Figuur 4. Figuratieve illustraties van: a. beelden gebruikt in gedragsexperiment; b. trainingsdata artificiële neurale netwerk (ANN); c. evaluatiedata ANN



actief is, terwijl de verbinding tussen S- en C-lagen exciterend is. Een dergelijk netwerk werd getraind om cijfers te herkennen van verschillende afmetingen en op verschillende posities (zie figuur 4b). Na training werden aan het netwerk afbeeldingen getoond van een groot cijfer opgebouwd uit de herhaling van een ander, kleiner cijfer (figuur 4c). Door de sterkte van de exciterende en remmende verbindingen te manipuleren, kon het netwerk ertoe gebracht worden het grote dan wel het kleine cijfer te herkennen. Dit weerspiegelt het contrast tussen neurotypische individuen (focus op globaal patroon) versus individuen met ASS (focus op lokaal patroon).

Beperkingen

ANN's kennen verschillende beperkingen als hersenmodel. Ten eerste ligt het aantal neuronen in een typisch hedendaags ANN veel lager dan in een menselijk brein. Toekomstige studies zullen moeten uitwijzen of het noodzakelijk is om het hele centrale zenuwstelsel artificieel te modelleren om zijn mechanismen te bestuderen.

Ten tweede is het puur wiskundige mechanisme waarmee een ANN typisch getraind wordt niet aanwezig in de hersenen. Ook hier is de vraag in welke mate het gebruik van een andere leermethode de validiteit van een model in de weg staat; hoe de concrete verbindingsterktes tussen neuronen precies tot stand zijn gekomen, is minder relevant dan de uiteindelijke netwerkconfiguratie zelf. Wel relevant is dat een ANN, eenmaal getraind, niet meer bijleert, in tegenstelling tot de hersenen. Sterker nog: wanneer een getraind ANN iets nieuws bijgeleerd wordt met extra training, bijv. een extra neuropsychiatrische aandoening om te classificeren, zal dit vaak resulteren in het 'vergeten' van eerder geleerde informatie, het zogenaamde *catastrophic forgetting*.[4] Een ander probleem is dat getrainde modellen doorgaans niet goed veralgemenen naar afwijkende data van die waarmee ze getraind zijn (bijv. een ongezien hondenras of bloemensoort). Erger nog, het is vaak mogelijk om een voor de mens onzichtbare verandering aan een input toe te voegen (bijv. wijzigingen aan enkele fotopixels) die zorgt dat het netwerk een specifiek *foutief* label genereert. Dit staat bekend als *adversarial attacks*.[4]

Conclusie

In dit artikel trachten wij met enkele concrete voorbeelden te schetsen wat artificiële neurale netwerken zijn en hoe ze gebruikt kunnen worden om mechanismen in de hersenen te modelleren. Ook beschreven we een aantal beperkingen van deze aanpak. De huidige kennis en rekenkracht maken nog geen model mogelijk dat de hele informatieverwerking in de hersenen modelleert, maar met de huidige technieken behaalt men al waardevolle resultaten wat betreft het uittesten van hypotheses aangaande neuropsychiatrische mechanismen.

Het potentieel van ANN's in de neuropsychiatrische praktijk zien we op verschillende gebieden. We denken daarbij aan beslissingsondersteunende hulpmiddelen in de diagnostiek, waarbij patronen in multimodale gegevens (psychiatrische, cognitieve, affectieve, wearables, etc.) via een objectief mechanisme geclassificeerd kunnen worden. Ook bij preventie en prognose kunnen het ontstaan en verloop van stoornissen en symptomen voorspeld worden. Ten slotte zouden ANN's een belangrijk hulpmiddel kunnen zijn voor de therapie, om *precision medicine* verder te ontwikkelen zodat de behandeling afgestemd kan worden op individuele patiëntenkenmerken.

LITERATUUR

1 Vandermeulen R, Morissette L, Chartier S. Modeling prosopagnosia using dynamic artificial neural networks. Proc. International Joint Conference on Neural Networks 2011; 2074-9.

2 Hoffman R, Mcglashan TH. Synaptic elimination, neurodevelopment, and the mechanism of hallucinated 'voices' in schizophrenia. Am J Psychiatry 1998; 154: 1683-9.

3 Nagai Y, Moriwaki T, Asada M. Influence of excitation/inhibition imbalance on local processing bias in autism spectrum disorder. Proceedings of the 37th Annual Meeting of the Cognitive Science Society, July 23-5, 2015. Pp. 1685-90.

4 Serre T. Deep learning: the good, the bad, and the ugly. Annu Rev Vision Sci 2019; 5: 399-426.