Hoe verder met AI en embedded systems?

tools, platforms – en worden schrijvers obsoleet?

Stuart Cording (Elektor)

AI wordt vaak gezien als een universeel probleemoplossend instrument dat elke uitdaging aankan. Maar ook wordt beweerd dat het het einde van de beschaving zal betekenen. Als technici weten we dat geen van beide gevallen waar is. Maar hoe kunnen we beter gebruik maken van AI, of eigenlijk machine learning, in de toepassingen die we ontwikkelen? En zullen de meest recente ontwikkelingen op het gebied van AI ons echt overbodig maken?

Het noemen van kunstmatige intelligentie (AI) trekt gegarandeerd de aandacht van de reguliere pers. Dankzij internet en cloud services, gecombineerd met soms twijfelachtige informatiebronnen, lijken overal nieuwe AI-gestuurde services op te duiken. Met een paar muisklikken kunnen je woorden in de mond van Morgan Freeman worden gelegd of kan je foto worden geïntegreerd in een nieuw kunstwerk. En hoewel dit amusant en tot nadenken stemmend kan zijn, is het niet gemakkelijk om een link terug te vinden naar de wereld van embedded systems. Maar dat is precies wat er gaande is binnen de industrie.

De meeste embedded systemen gebruiken op regels gebaseerde programmeermethoden om hun functionaliteit te verwezenlijken. Op basis van de input bepaalt een reeks if/else- of switch-statements hoe er gereageerd moet worden. Dit werkt goed voor een beperkt aantal ingangen. Maar op een gegeven moment maakt het aantal ingangen of de complexiteit van hun relatie het moeilijk om duidelijke programmeerregels te definiëren.

Stel je bijvoorbeeld een motor voor in een fabriek die 24 uur per dag, zeven dagen per week draait. De ervaring leert dat na verloop van tijd slijtage optreedt en het lagervet dikker wordt. Mettertijd veranderen hierdoor de opstarttijd, het geproduceerde geluid, de trillingen, de bedrijfstemperatuur van de motor en de opgenomen stroom. Regelmatig onderhoud is de manier waarop deze problematiek tegenwoordig wordt aangepakt, maar het resulteert in regelmatige onderbrekingen die de productie stilleggen. Of dit te vaak of niet vaak genoeg gebeurt is vaak moeilijk te zeggen. Bovendien is het onwaarschijnlijk dat een potentiele storing als gevolg van een haarscheurtje in de as, lagers, behuizing of bevestigingen wordt opgemerkt. Mits goed getraind kunnen AI-methoden potentiële storingen vaststellen aan de hand van een complexe mix van informatiebronnen. Als dergelijke intelligentie kan worden toegepast in een systeem op basis van een microcontroller, krijg je een betaalbaar controlesysteem dat tijd en geld bespaart, en verspilling door onnodige smering en vervanging van onderdelen vermindert.

Microcontrollers intelligent maken

Op het niveau van de microcontroller hebben we het eerder over machine learning (ML) dan over AI. Dit houdt in dat een machine geprogrammeerd wordt om regels te gebruiken die ontwikkeld zijn om op basis van de analyse van beschikbare gegevens beslissingen te nemen. Hoewel microcontrollers meer dan krachtig genoeg zijn om dergelijke ML-algoritmen uit te voeren, blijft het leren van trainingsgegevens buiten hun bereik en is daarvoor minstens een

desktopcomputer nodig, en misschien zelfs een cloudserver. Edge Impulse, opgericht in 2019, heeft een platform ontwikkeld dat gewijd is aan ML in embedded systems, en heeft met succes samengewerkt met halfgeleiderleveranciers wereldwijd [1] om brede ondersteuning uit te rollen.

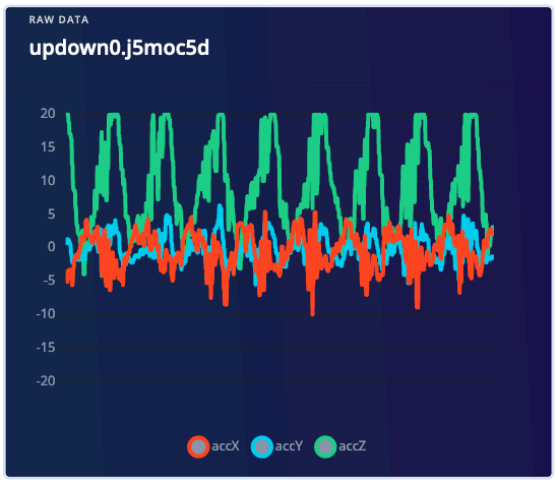
Het uitgangspunt voor elke ML-toepassing is data. Terwijl sommige toepassingen, zoals autonoom rijden, terabytes aan trainingsdata nodig hebben, kunnen eenvoudige microcontroller-gebaseerde systemen leren van slechts een paar kilobytes aan data. De gegevens uit het board in de Edge Impulse-omgeving te krijgen, is dus de eerste uitdaging. De eerste gedachte zou zijn om de seriële interface van een Arduino te gebruiken om de gegevens naar de PC te sturen en ze van daaruit als tekstbestand door te sturen. Hun platform is echter ingesteld om de gegevens rechtstreeks te verwerken.

Data – voer voor AI

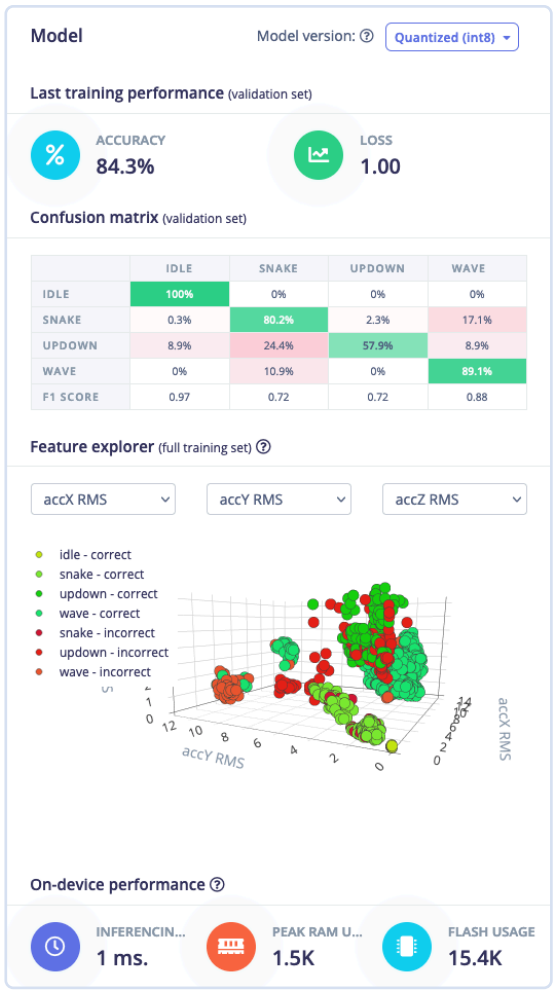
Eén hulpmiddel is de Data Forwarder [2], een commandoregel-toepassing (CLI) die gegevens van een ontwikkelboard rechtstreeks naar de Edge Impulse-omgeving stuurt. Met je gebruikersnaam en wachtwoord wordt een verbinding tot stand gebracht tussen de seriële poort op je PC en de server. Aan de kant van de microcontroller is uitvoer van data via de seriële interface in een komma- of tabgescheiden formaat alles wat nodig is. Zolang de bemonsteringsfrequentie relatief laag is, is dit een ideale manier om rechtstreeks representatieve gegevens van je sensoren te verzamelen (**figuur 1**). Krachtiger embedded systems, zoals de Raspberry Pi of de NVIDIA Jetson Nano, kunnen gebruik maken van de meegeleverde software development kit (SDK) [3]. Die ondersteunt ook sensoren zoals microfoons en camera's die grotere hoeveelheden data genereren.

Met de data die naar Edge Impulse zijn verstuurd, is de volgende stap het definiëren van een 'impuls'. Deze bestaat uit twee blokken. De eerste hakt de data in kleinere brokken en gebruikt signaalverwerkingstechnieken om er informatie uit te halen. Dit zorgt ervoor dat de beschikbare sensorgegevens worden omgezet in consistente informatie voor de tweede verwerkingsfase. Het tweede blok is waar het leren en classificeren plaatsvindt (**figuur 2**). In een voorbeeldprojectmet de naam "Continuous Motion Recognition" wordt goed uitgelegd hoe deze blokken worden geconfigureerd om accelerometerdata te analyseren en deze input te classificeren als een van vier gebaren [4].

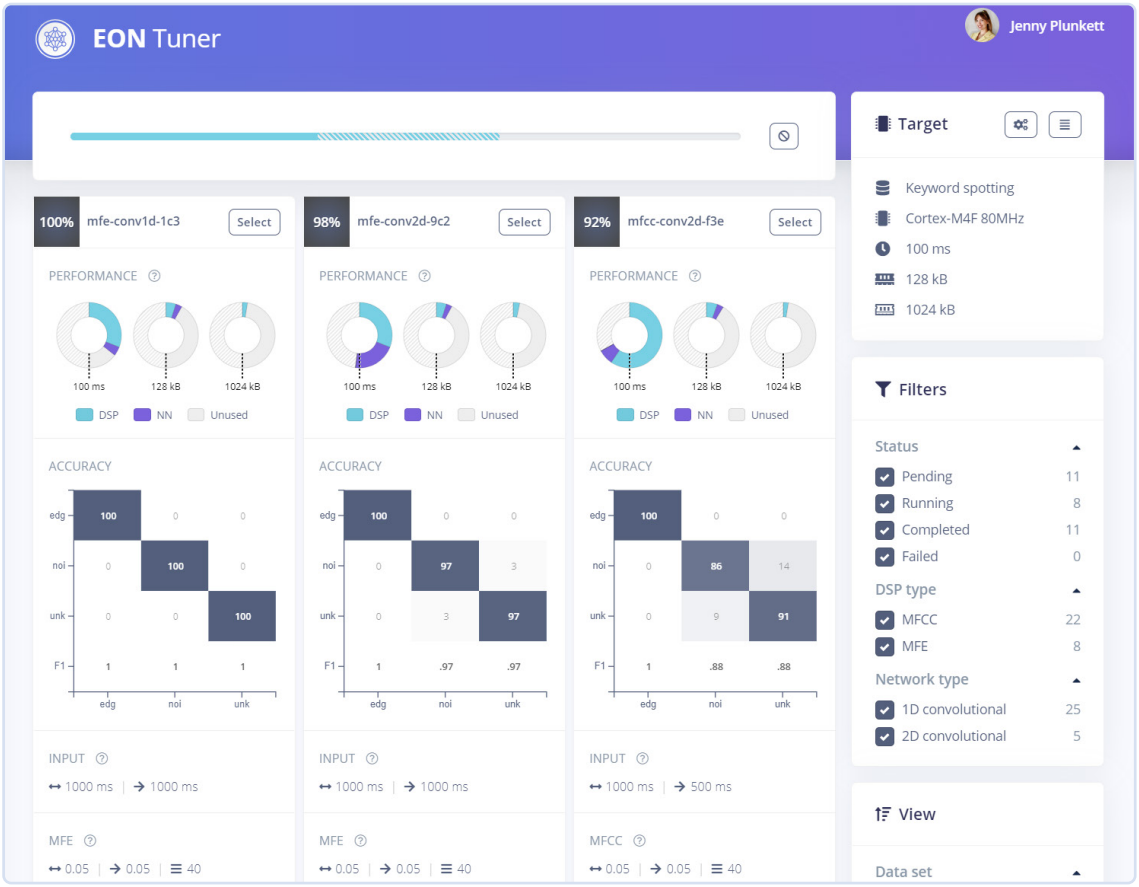
Dit is de meest kritische stap in elke ML-ontwikkeling, waarbij vaak creatief moet worden nagedacht over de beste aanpak. Soms is het negeren van de invoer van sommige sensoren de beste oplossing, terwijl in andere gevallen juist meer gegevens nodig zijn. Misschien merk je dat je teveel leert, of dat het gekozen neurale netwerkmodel slecht past bij de classificatie die je probeert te maken. Een andere cruciale stap is de classificatie van afwijkingen. In het voorbeeldproject zijn er vier gedefinieerde gebaren. Andere bewegingen die op de geleerde gebaren lijken, moeten echter worden uitgesloten. Goede anomaliedetectie levert een betrouwbaarder ML-resultaat op.



Figuur 1. Een drieassige accelerometer levert bewegingsdata om een toepassing te ontwikkelen voor ML-gebarenherkenning (bron: Edge Impulse).



Figuur 2. De Edge Impulse-omgeving geeft feedback over nauwkeurigheid, snelheid en geheugengebruik na een eerste trainingscyclus (bron: Edge Impulse).



Figuur 3. ML-algoritmen worden verder geoptimaliseerd voor het doelsysteem met behulp van EON Tuner, waardoor de inferentie-responstijden verbeteren en het geheugengebruik afneemt (bron: Edge Impulse).

De laatste stap is de implementatie. Via de webinterface wordt firmware voor het gekozen apparaat gedownload voor integratie in je toepassing. Voor Arduino wordt een bibliotheek gegenereerd, terwijl je voor andere microcontrollers een C++ bestand kunt genereren. Natuurlijk variëren de prestaties van microcontrollers enorm. Om het best mogelijke resultaat te garanderen, biedt Edge Impulse de EON Tuner [5] aan. Dit tool kan de detectienauwkeurigheid verder verbeteren, de inferentie versnellen en de geheugenvereisten verminderen door informatie te gebruiken zoals doelapparaat, grootte van het geheugen en latentie (**figuur 3**).

Figuur 4. SlateSafety gebruikte Edge Impulse om edge ML fysiologische monitoring en de BAND V2 veiligheidsmonitoring te verbeteren (bron: SafetySlate).

ML in echte toepassingen

Echte toepassingen gebruiken deze aanpak om ML-functionaliteit in te bouwen. BAND van SlateSafety [6] integreert een reeks biometrische sensoren om werknemers te controleren die onder moeilijke omstandigheden werken (**figuur 4**). Deze variëren van eerstehulpverleners tot werknemers in de industrie die zware persoonlijke beschermingsuitrusting dragen, zoals brandweerlieden. Die uitrusting stuurt gewoonlijk gegevens naar de cloud, zodat collega's de vitale functies van een gebruiker kunnen controleren. Maar vooral bij rampen kunnen verbindingen haperen of compleet uitvallen. Het ontwikkelingsteam gebruikte Edge Impulse om edge ML te integreren in de bestaande uitrusting, getraind op historische biometrische gegevens. Met behulp van de EON Tuner werd het algoritme optimaal afgestemd op de hardware en vervolgens ingezet via een over-the-air update. Nu kan BAND de drager een waarschuwing geven wanneer er een risico op een hittecollaps bestaat, zelfs wanneer er geen draadloze verbinding is.

Productontwikkeling verbeteren met AI

Natuurlijk hoeft AI niet geïntegreerd te zijn in het product. Het kan ook voor de ontwikkeling ervan worden gebruikt. Tegenwoordig worden veel complexe toepassingen op basis van modellen ontwikkeld, waarbij in wezen een beschrijving wordt ontwikkeld van hoe iets zou werken met behulp van software en

wiskundige vergelijkingen die hun oorsprong vinden in de fysica. Maar zelfs deze aanpak heeft zijn grenzen. Dit is waar Monolith en zijn zelflerende AI-platform [6] om de hoek komen kijken.
Het platform is in staat de fysieke eigenschappen van complexe systemen te leren op basis van reeds verzamelde gegevens. Voertuigen doorlopen bijvoorbeeld een serie tests op een testparcours, waarbij meerdere sensoren slinger- en -draaibewegingen monitoren, samen met de wielsnelheid en de acceleratie. Het verzamelen van gegevens voor verschillende stijfheden van de ophanging geeft een goed inzicht in hoe het voertuig reageert op verschillende rijsituaties. Normaliter worden de gegevens geanalyseerd en vervolgens in nieuwe instellingen voor de volgende testrit gebruikt. Monolith kan de gegevens van de eerste reeks testritten evalueren en het resultaat van veranderingen aan de ophanging met grote nauwkeurigheid voorspellen. De resultaten kunnen worden gebruikt om sneller de beste instellingen voor de vering te bepalen, zodat er minder extra fysieke testritten nodig zijn.
Deze benadering kan ook worden toegepast op metrologie. Gasmeters moeten uitzonderlijk nauwkeurig zijn om een correcte facturering te garanderen, maar dat vormt een uitdaging wanneer de meter een reeks verschillende gassen moet meten. Bij een klant kwam bij het simuleren van ultrasone meters een puur wiskundige analyse aan de grens van het mogelijke, zodat voor de kalibratie slechts herhaalde testprocessen overbleven als om de benodigde certificering te behalen. Gelukkig leverde al die tests een rijke verzameling gegevens op voor analyse. Met behulp van zelflerende AI-modellen is de hoeveelheid benodigde tests met maar liefst 70% afgenomen, waardoor de ontwikkeling aanzienlijk is versneld.

Forse verkleining van AI-computers

Wedstrijden zoals de DARPA Grand Challenge [8], waarbij teams autonome voertuigen bouwden die een bochtig parcours moesten kunnen afleggen, veroorzaakten een golf van belangstelling voor zelfrijdende auto's. Nu, bijna twintig jaar later, is er veel geld uitgegeven maar is daar weinig uit voortgekomen. Tesla haalt op dit gebied vaak het nieuws, vooral wanneer overenthousiaste Tesla-bezitters te veel vertrouwen stellen in de mogelijkheden van hun voertuig [9]. Voor het overige lijkt alleen Waymo echte zelfrijdende voertuigen op afroep aan te bieden [10], maar die rijden uitsluitend in Phoenix en San Francisco in de VS.
Een van de problemen is dat een computer die een auto bestuurt een uitzonderlijke uitdaging vormt. Niet alleen moet het voertuig voortdurend de situatie om zich heen beoordelen, maar het moet ook de acties van andere bestuurders en weggebruikers zoals voetgangers en fietsers voorspellen, die zich misschien niet aan de verkeersregels houden.

Figuur 5. De ontwikkeling van AI-algoritmen wordt versneld met behulp van krachtige in het voertuig geplaatste computers, zoals deze vloeistofgekoelde DYNACOR 40-36 (bron: Eurotech).

Wat er gebeurt is dat de elektrische en elektronische (E/E) architectuur van voertuigen verandert om te voldoen aan de toekomstige behoeften van autonome voertuigen. Met een veelheid aan sensoren die enorme hoeveelheden gegevens opleveren, gaat de industrie over op automotive Ethernet. Momenteel vormt deze benadering de basis van geavanceerde rijhulpsystemen (ADAS) die, door de controle van remmen, versnelling en stuurinrichting, kunnen ingrijpen als wij bestuurders een fout maken. Volgens de autonomieniveaus van de Society of Automotive Engineers (SAE) voldoen auto's uit de topklasse momenteel aan niveau 2+, waarbij sommige aan niveau 3 tippen. Volledige autonomie waarbij we achterover kunnen leunen en ons kunnen ontspannen is echter niveau 5, dus daar moet nog veel gebeuren.
Bedrijven als Eurotech ondersteunen de industrie om de ontwikkeling van de benodigde algoritmen te versnellen. Momenteel verzamelt een acht uur durende testrit 120 TB aan gegevens die terug moeten naar het lab voor verwerking en analyse. Verbeteringen van AI-algoritmen kunnen in het laboratorium worden getest met behulp van de verzamelde gegevens, maar er is weinig beschikbaar om het testen en de ontwikkeling van algoritmen in het veld te ondersteunen.
Gebruikmakend van zijn ervaring met vloeistofkoeling biedt Eurotech een serie geavanceerde AI-hardware die daarvoor geschikt is – in wezen compacte supercomputers die in de kofferbak van een voertuig passen. Een apparaat zoals de DYNACOR 40-36 is robuust opgebouwd voor gebruik in zowel on- als off-road voertuigen [11]. Voorzien van een 16-core Intel Xeon CPU met 64 GB RAM en maximaal twee NVIDIA GV100 GPU's met 32 GB RAM, levert deze ventilatorloze computer 237 TFLOPS om deep learning-toepassingen aan te kunnen (**figuur 5**). Verschillende gigabit Ethernet-interfaces ondersteunen het inlezen van de enorme hoeveelheden sensorgegevens, van radar en camera's tot lidar, en sluizen die naar een 32 TB solid-state geheugen door. Wanneer tijdens testritten meer

Figuur 6: Het door DALL-E 2 AI van OpenAI gegenereerde beeld van een zelfrijdend voertuig onder besturing van de vertrouwde Commodore 64. Een werk in uitvoering.

inferentie- en versterkingstests kunnen worden uitgevoerd, zou de weg naar niveau5-autonomie aanzienlijk versneld kunnen worden.

Brengt AI mijn baan in gevaar?

Een voortdurende discussie op sociale media is of de vooruitgang in AI banen in creatieve bedrijfstakken in gevaar brengt. De lancering van DALL-E 2 door OpenAI [12] zet gesproken opdrachten om in afbeeldingen (**figuur 6**). Maar misschien nog indrukwekkender is zijn vermogen om bestaande beelden realistisch te bewerken. Het kan bijvoorbeeld objecten in de voor- of achtergrond verwijderen. En op basis van een werk van de Nederlandse schilder Vermeer, kan de AI zijn "Meisje met de parel" uitbreiden met een geloofwaardige impressie van de kamer waarin ze zat toen ze werd geschilderd.

Schrijvers, zoals de redacteuren van Elektor en andere gerenommeerde uitgeverijen, zijn echter geschokt door de lancering van ChatGPT [13]. Deze AI kan met gebruikers in spreektaal communiceren in een veelheid van talen. Tot nu toe hebben discussies over de voordelen van siliciumcarbide MOSFET's (SiC) en galliumnitride transistoren (GaN) ten opzichte van silicium MOSFET's een zeer nauwkeurige respons opgeleverd. Dus zelfs niche-onderwerpen lijken goed te worden bestreken.

Hoewel uitzonderlijk slim, kent het hulpmiddel alleen antwoorden op onderwerpen die zich voordeden voordat het werd getraind. Omdat het niet voortdurend bijleert, zal het niet op de hoogte zijn van actuele zaken of de nieuwste drama's rond K-pop bands (jammer). Een ander kritiekpunt is dat de antwoorden na een tijdje ietwat gekunsteld en formeel lijken. Maar wie op zoek is naar inspiratie of een verjaardagsrijm in de stijl van een beroemd dichter, zal niet teleurgesteld worden. Om te bepalen of de toekomst van Elektor afhankelijk is van wezens van vlees en bloed of van computers, geef ik je een samenvatting van het onderwerp van dit artikel geschreven door ChatGPT. Tot de volgende keer... of misschien toch niet!

Samengevat zijn embedded systems en AI twee technologieën die steeds vaker samen worden gebruikt om intelligente, autonome apparaten en systemen te maken. Embedded systemen leveren het hardware- en softwareplatform waarop AI-algoritmen kunnen draaien, terwijl AI-algoritmen deze systemen in staat stellen hun omgeving op een intelligentere en menselijkere manier waar te nemen, te analyseren en erop te reageren. Naarmate de mogelijkheden van zowel embedded systemen als AI verder verbeteren, kunnen we een breed scala aan spannende nieuwe toepassingen verwachten op gebieden als robotica, gezondheidszorg, vervoer en meer.

220673-03

Vragen of opmerkingen?

Hebt u technische vragen of opmerkingen naar aanleiding van dit artikel? Stuur een e-mail naar de auteur via stuart.cording@elektor.com of naar de redactie van Elektor via redactie@elektor.com.

WEBLINKS

[1] Edge Impulse partners: http://bit.ly/3vbqRhG
[2] Edge Impulse CLI installatie: http://bit.ly/3BQQt71
[3] Edge Impulse Ingestion SDK: http://bit.ly/3jplhp3
[4] Continuous motion recognition voorbeeldproject: http://bit.ly/3WAV9G5
[5] EON Tuner: http://bit.ly/3PQhFsr
[6] SlateSafety BAND: http://bit.ly/3YVpe5x
[7] Monolith: http://bit.ly/3BWZImh
[8] DARPA Grand Challenge, Wikipedia: http://bit.ly/3VnqWJI
[9] J. Stilgoe, "Tesla crash report blames human error – this is a missed chance," Guardian, januari 2017: http://bit.ly/3Vjp7gI
[10] DYNACOR 40-36: http://bit.ly/3GdmgBS
[11] Waymo One: http://bit.ly/3FNG8Ku
[12] DALL·E 2: http://bit.ly/3PNPWsD
[13] ChatGPT: http://bit.ly/3PLjZRo