

Een introductie tot TinyML

Mark Patrick (Mouser Electronics)

Dit artikel onderzoekt concepten van machine learning (machinaal leren) met behulp van eenvoudige, low-power microcontrollers, onder de noemer TinyML. Machine learning blijft zijn stempel drukken op vele aspecten van ons dagelijks leven, thuis, op kantoor of waar dan ook. Hoewel veel applicaties voor machine learning aanzienlijke rekenkracht vereisen om complexe wetenschappelijke of financiële gegevens te kraken, bieden ze slechts beperkte reken- en verbindingsmogelijkheden voor het internet der dingen (IoT) en andere randtoepassingen.

AI en machinaal leren: integraal onderdeel van ons leven

Praten tegen ons polshorloge of een Star Trek communicator-gadget was nog niet zo lang geleden science fiction, maar nu is dat precies wat velen van ons dagelijks doen. We praten met onze smartphone-apps, met het infotainment-systeem van onze auto's en onze slimme "smart speakers" thuis.

Kunstmatige intelligentie (AI) en het creëren van zelfdenkende computers die voelen, herkennen en problemen oplossen, vormen nu de basis van de hedendaagse computer- en datawetenschap. Machine learning (ML) is een toepassingsgebied van AI en richt zich op het gebruik van algoritmen die computers in staat stellen te leren hoe ze een taak beter moeten uitvoeren zonder expliciet daarvoor te zijn geprogrammeerd.

Een groot deel van de dagelijkse realiteit wordt al vormgegeven door machine learning, van weersvoorspellingen en het vinden van de beste route voor het

woon-werkverkeer tot de reclamebanners en advertenties op je favoriete social media-app. Veel gebieden van wetenschappelijk onderzoek zijn nu afhankelijk van ML om petabytes aan gegevens te doorzoeken op trends.

Machine learning – hoe werkt het?

Van de hierboven genoemde voorbeelden van machine learning zien we meestal alleen de resultaten. We hebben vaak geen idee van de complexiteit achter het vinden van een nieuw zwart gat bijvoorbeeld, of het voorspellen van het weer aan de hand van het aantal permutaties van eerdere weersomstandigheden. Veel complexe taken vereisen grote datasets, meerdere algoritmen en een aanzienlijke rekenkracht. Als je je verder verdiept in machine learning, vind je verschillende categorieën van leren, elke geschikt voor een specifieke taak. We gaan er hier niet dieper op in, maar ze worden onderverdeeld in "supervised", "unsupervised", "semi-supervised" en "reinforcement learning". Voor elke ML-toepassing is een neurale netwerk cruciaal. Simpel gezegd, een neurale netwerk is een poging om de functie van de neuronen in het menselijk brein te repliceren in een wiskundig model. Dit model maakt gebruik van algoritmen om de waarschijnlijkheid van een resultaat af te leiden. Bij beeldherkenning bijvoorbeeld is er 95% waarschijnlijkheid dat de afbeelding op de foto een hond is. Er zijn verschillende soorten neurale netwerken en je komt termen tegen als convolutionele neurale netwerken (CNN) en terugkerende neurale netwerken (RNN). Elk soort neurale netwerk heeft een andere set onderling verbonden lagen, waardoor de één beter geschikt is voor een specifieke taak dan de ander.

Bij supervised learning wordt het neurale netwerk gevoed met trainingsgegevens, zodat een algoritme een resultaat kan afleiden. Voor beeldherkenning bijvoorbeeld is een CNN het beste. Om verschillende soorten fruit te herkennen, moet je het neurale netwerk-algoritme voeden met duizenden gelabelde

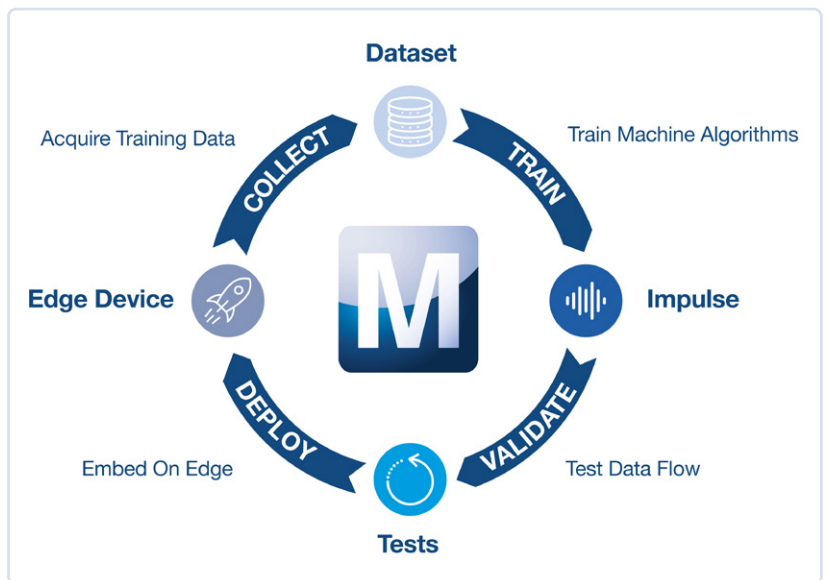
beelden van verschillende soorten fruit met diverse gradaties van rijpheid, gefotografeerd vanuit verschillende hoeken. Het algoritme zoekt dan naar herkenbare kenmerken waarmee het de diverse typen fruit van elkaar kan onderscheiden. De trainingsfase is iteratief en kan inhouden dat het algoritme, wanneer het wordt geconfronteerd met een testdataset van beelden, wordt bijgesteld om de hoogste waarschijnlijkheid te bereiken.

Zodra de testgegevens de beste algoritme-prestaties van het neurale netwerk opleveren, is het model klaar voor gebruik. Bij inzet, ook bekend als inferentie, leidt het model resultaten af op basis van waarschijnlijkheid. Onze dagelijkse interactie met bijvoorbeeld onze slimme speaker omvat meestal een triggerwoord of -zin, zoals "Hé, Google" om hem te wekken, zodat hij begint te luisteren naar wat er gaat komen. Een slimme speaker beschikt niet over de rekencapaciteit van een datacentrum, dus worden er korte audiobestanden opgenomen en naar de cloud gestreamd voor inferentie om de aard van ons verzoek te bepalen. Het detecteren van het triggerwoord of de triggerzin is een uitstekend voorbeeld van eenvoudig machine learning. Welkom bij TinyML!

Machine learning en het industriële IoT

Nu machine learning alomtegenwoordig wordt, groeit de potentiële lijst van applicaties exponentieel. Veel industriële toepassingen zijn bijvoorbeeld niet gericht op "big data", maar op de vraag hoe we productielijnen efficiënter kunnen maken. De kosten van een onverwachte storing in een productieproces kunnen stevig oplopen. Een motorstoring tijdens een gekoeld mengproces van voedingsmiddelen kan de productie stilleggen, waardoor er grondstoffen verloren gaan. Veel organisaties passen nu predictief onderhoud toe door een stilstand te plannen om dergelijke storingen te voorkomen. Toestandsafhankelijk monitoren van productiemiddelen, zoals motoren en aandrijvingen, helpt bij het voorspellen wanneer bijvoorbeeld een motorlager tekenen van overmatige slijtage gaat vertonen. Machine learning-algoritmen in IIoT-edgesensoren kunnen vaststellen wanneer de trillings-signatuur van een motor afwijkt van de norm en zo tijdig waarschuwen.

Alleen al binnen het industriële domein zijn de toepassingen voor machine learning enorm, maar deze gaan ook gepaard met verschillende technische uitdagingen. In tegenstelling tot de "big data"-toepassingen zijn de reken- en geheugenbronnen van een eenvoudige IIoT-edgesensor slechts een fractie van die in een datacenter. Een fabriek kan honderden sensoren hebben, dus schaalvoordeel is een factor, evenals de fysieke grootte, de beschikbaarheid van een geschikte stroomvoorziening en bekabelde of draadloze connectiviteit. In de meeste gevallen is het neurale netwerk-



Figuur 1. Training en inferentie.

algoritme van een trillingssensor minder veeleisend dan bijvoorbeeld sterrenkundig onderzoek, dus zoeken embedded ontwikkelaars naar manieren om netwerkmodellen te draaien op low-power microcontrollers met batterijvoeding.

TinyML: machine learning aan de edge

Het voor de training van een neurale netwerkmodel gebruikte computerplatform hoeft niet hetzelfde te zijn als het platform dat wordt ingezet, waardoor een van de meest hulpmiddel-intensieve processen wegvalt. Er zijn echter nog steeds technische uitdagingen. Kan het algoritme tijdig worden uitgevoerd? Hoe communiceert het apparaat met een hostsysteem om het operationele personeel te informeren? Ook de embedded ontwikkelaars worden met uitdagingen geconfronteerd. De meesten van hen zijn geen datawetenschappers, dus het leren van de concepten van machine learning en het werken met neurale netwerken vormen een steile leercurve. [▶](#)

230062-03



Over de auteur

Als Technical Marketing Manager voor EMEA bij Mouser Electronics is Mark Patrick verantwoordelijk voor het maken en verspreiden van technische content in die regio – content die essentieel is voor de strategie van Mouser om zijn technische publiek te ondersteunen, te informeren en te inspireren. Voordat hij leiding gaf aan het Technical Marketing-team, maakte Patrick deel uit van het EMEA Supplier Marketing-team en speelde hij een cruciale rol bij het aangaan en ontwikkelen van relaties met belangrijke productiepartners. Naast een verscheidenheid aan technische en marketingfuncties, heeft Patrick onder andere acht jaar bij Texas Instruments in Applications Support en Technical Sales gewerkt. Als 'hands-on' engineer in hart en nieren, met een passie voor vintage synthesizers en motorfietsen, draait hij zijn hand er niet voor om die te repareren. Patrick heeft een eerstegraads Honours Degree in Electronics Engineering van Coventry University.