

Embedded machine learning voor microcontrollers

16 januari 2024, 16:17

Frank Van den Broek

Embedded machine learning in praktijk: computervisie op een Raspberry Pi Pico microcontroller

Tot voor kort was het domein van edge machine learning (edge ML) bijna uitsluitend gericht op inferentie op vaste computers met GPU's of op mobiele apparaten. Met de opkomst van tiny machine learning (TinyML) zullen neurale netwerken en data-analyses ook toegepast kunnen worden op microcontrollers. Om de mogelijkheden van TinyML te demonstreren, heeft Sirris een computervisietoepassing uitgerold op een goedkope microcontroller: de Raspberry Pi Pico.

TinyML is een snelgroeiende multidisciplinaire technologie waarbij innovaties op het gebied van embedded hardware, software en machine learning een nieuwe klasse slimme toepassingen mogelijk maken door middel van on-device, near-sensorinferentie op ultra-low power embedded systemen. Dit opent de deur naar nieuwe soorten edge services en toepassingen die niet afhankelijk zijn van cloud processing, maar gebaseerd op inferentie on the edge.

Omdat data on-device worden verwerkt en niet verstuurd via een netwerk, biedt TinyML grote voordelen op het gebied van privacy, latency, energie-efficiëntie en betrouwbaarheid. Een bekende eerste toepassing van TinyML op always-on apparaten die op batterijen werken, was de zogenaamde audio-wake-word-detectie op producten van tech-giganten ("Okay Google", "Hey

Siri"). Dankzij recente ontwikkelingen op het gebied van TinyML-ontwikkelplatformen, machine learning modellen en embedded hardware, ligt embedded machine learning nu ook binnen het bereik van de bredere industrie voor veel verschillende toepassingen met sensoren zoals camera's, versnellingsmeters en geluidssensoren.

Computervisie op een goedkope microcontroller

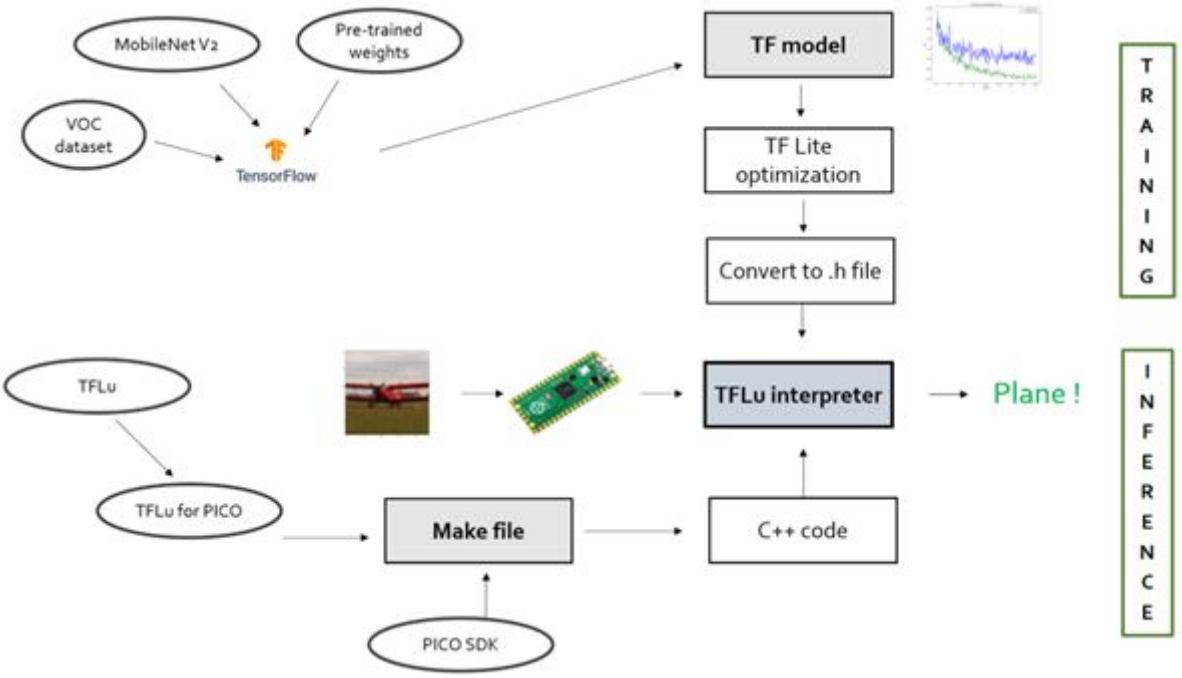
Om de mogelijkheden van TinyML te demonstreren, heeft Sirris met open-source-tools een computervisietoepassing uitgerold op een goedkope microcontroller: de Raspberry Pi Pico (kostprijs: minder dan 5 euro). Het doel was de haalbare prestaties te beoordelen, waaronder nauwkeurigheid en inferentietijd voor een beeldclassificatietaak, met behulp van open-source-computervisiemodellen en -tools.

Computervisie heeft het afgelopen decennium een grote doorbraak gekend dankzij deep learning algoritmen, AI-netwerken die vroeger veel geheugen en rekenkracht nodig hadden en daarom niet konden worden toegepast op microcontrollers. De Raspberry Pi Pico heeft bijvoorbeeld slechts 264 KB SRAM / 2 MB flashgeheugen en een 133 MHz Arm Cortex-M0+ processor. Nu bestaan er, specifiek voor mobiele en embedded apparaten, nieuwe deep learning netwerkarchitecturen die goed presteren voor computervisietaken, zelfs met beperkte rekenkracht. In ons geval hebben we een MobileNetV2 gebruikt, een voorgetraind convolutioneel neuraal netwerk dat vrij beschikbaar is en dat we voor onze toepassing verder hebben getraind op de VOC-2012 dataset via transfer learning.

Open-source-tools

Het trainen van een TinyML-model gebeurt met een vergelijkbaar trainingsproces als voor grotere deep-learning-modellen. Voor onze toepassing gebruikten we Tensorflow, een open-source ML framework van Google. Het resultaat was een getraind model klaar om in productie te gaan op een desktop- of serverplatform.

Maar om dit model op een microcontroller te implementeren, zijn verdere optimalisatiestappen nodig om het benodigde geheugen en de benodigde rekenkracht te verlagen. Het gaat dan bijvoorbeeld om zogenaamde ‘kwantisatie’, waarbij inputs en parameters met minder precisie worden weergegeven, en ‘pruning’, waarbij neuronen die weinig invloed hebben op het eindresultaat uit het netwerk worden verwijderd. Voor deze optimalisaties gebruikten we TensorFlow Lite, een andere open-source-tool van Google, die we vervolgens converteerden naar een C/C++ headerbestand en konden flashen op de Raspberry Pi Pico. Om het model op de microcontroller te gebruiken, gebruikten we de C/C++ bibliotheek TensorFlow Lite voor Microcontrollers (TFLu) van Google.



©Sirris

Resultaten

We konden probleemloos een computervisietoepassing implementeren op een Raspberry Pi Pico met opensourcetools. Het model verwerkt RGB-afbeeldingen met een resolutie van 48 x 48 en classificeert deze in drie klassen, met een inferentietijd van 250 ms.

In de toekomst zullen we de mogelijkheden verkennen van krachtiger microcontrollers, met meer geheugen en rekenkracht dan de Raspberry Pi Pico, op het gebied van onder andere computervisie en anomaliedetectie.

Webinar

Op 31 januari 2024 gaan we dieper in op de hierboven beschreven beeldclassificatietaak op een Raspberry Pi Pico microcontroller tijdens een gratis, één uur durend webinar. Tijdens dit webinar komen de volgende onderwerpen aan bod:

- Bespreken van de hardware, microcontroller en camera;
- Bespreken van de toegepaste MobileNet-architectuur en vergelijking met alternatieve netwerken;
- Een demonstratie van de opensourcetools die gebruikt worden voor training, optimalisatie en inferentie: TensorFlow, TensorFlow Lite, TensorFlow Lite voor Microcontrollers
- Bespreken van de resultaten.

Zin om erbij te zijn? Schrijf u dan nu in voor ons webinar!

Dit artikel is geschreven in het kader van ons CORNET COOCK-project EmbedML, gefinancierd door VLAIO.

Auteurs



Frank Van den Broek