KU LEUVEN

FACULTEIT INDUSTRIELE
INGENIEURSWETENSCHAPPEN

TECHNOLOGIECAMPUS GENT

Comparative Study of Single Board Computers for AI Edge Computing Purposes

Arno PLAETINCK

Promotor: Prof. Dr. Ir. Nobby Stevens

Co-promotoren: Ing. Willem Raes
Ing. Jorik De Bruycker

Masterproef ingediend tot het behalen van de graad van master of Science in de industriële wetenschappen: Industriële Ingenieurswetenschappen Elektronica-ICT Embedded Systems

Academiejaar 2019 - 2020

Dankwoord

In deze thesis wordt een benchmark voorgesteld die de performantie van verscheidene edge devices analyseert op vlak van latency. De uitvoering van verschillende getrainde neurale netwerken wordt vergeleken met parameters zoals processorgebruik, klokfrequentie, prijs en energieverbruik. De verworven inzichten kunnen toegepast worden in allerlei low-latency applicaties zoals zelfrijdende voertuigen.
Verder had ik graag de mensen bedankt die geholpen hebben deze thesis tot stand te brengen. Hulp van personen zoals mijn promotor Prof. Dr. Ir. Nobby Stevens en de co-promotoren Ing Willem Raes en Ing Jorik De Bruycker werd zeer geapprecieerd.

Abstract

Benchmark voor latency bij edge-devices.

Er is een hele grote verscheidenheid aan onderwerpen waarbij de duur van verkrijgen van resultaten binnen zekere tijdslimieten moeten vallen. Hiervoor wordt vaak een afweging gemaakt tussen cloud en edge computing. In de cloud kan met krachtige en snelle hardware worden gewerkt. Echter treedt hierbij een significante vertraging op, veroorzaakt door het verzenden van data over het internet. Uitvoeren van programma's in de edge hebben geen last van deze vertraging. Dit kan voordelig zijn bij latency-gevoelige applicaties zoals het zelfstandig rijden van voertuigen.

Deze scriptie bespreekt een benchmark van drie verschillende edge-toestellen en een reguliere computer. Deze benchmark biedt informatie over de latency en processor-performantie van elk device. Deze zal meer inzicht bieden wanneer men een kosten-baten analyse maakt. De onderzochte devices binnen deze thesis zijn de toestellen zoals de Google Coral Dev die over zo wel een CPU als TPU bezit en de Raspberry Pi die enkel over een CPU beschikt. Ook wordt een Nvidia Jetson Nano en een Personal Computer gebruikt die zowel een CPU als een GPU bezitten.

Om een representatieve benchmark te bekomen is het nodig om verschillende programma's op een identieke wijze te testen. De te beproeven programma's zullen verdeeld worden over de categorieën regressie en classificatie. De data die uit de benchmark afgeleid wordt bestaat uit de latency en het processor-gebruik van het runnen van elk programma. Deze worden dan herwerkt en genormaliseerd om de resultaten eenvoudiger te analyseren. Hierbij worden ook factoren zoals prijs, kloksnelheid en energieverbruik in rekening gebracht.

Uit de resultaten kan er een duidelijk verschil afgeleid worden tussen de verschillende edge-toestellen. Hierbij komt de Nvidia Jetson Nano het best naar voor. Het toestel behaalt de laagste latency voor alle toegepaste programma's. De resultaten geven ook aan dat de Coral Dev efficiënter met energie omspringt. Voor het uitvoeren van hetzelfde programma zal de Coral Dev minder energie verbruiken.

Trefwoorden: benchmark, edge-devices, latency, Jetson Nano, Coral Dev, Raspberry Pi

Abstract English

Benchmark voor latency bij edge-devices.

Er is een hele grote verscheidenheid aan onderwerpen waarbij de duur van verkrijgen van resultaten binnen zekere tijdslimieten moeten vallen. Hiervoor wordt vaak een afweging gemaakt tussen cloud en edge computing. In de cloud kan met krachtige en snelle hardware worden gewerkt. Echter treedt hierbij een significante vertraging op, veroorzaakt door het verzenden van data over het internet. Uitvoeren van programma's in de edge hebben geen last van deze vertraging. Dit kan voordelig zijn bij latency-gevoelige applicaties zoals het zelfstandig rijden van voertuigen.

Deze scriptie bespreekt een benchmark van drie verschillende edge-toestellen en een reguliere computer. Deze benchmark biedt informatie over de latency en processor-performantie van elk device. Deze zal meer inzicht bieden wanneer men een kosten-baten analyse maakt. De onderzochte devices binnen deze thesis zijn de toestellen zoals de Google Coral Dev die over zo wel een CPU als TPU bezit en de Raspberry Pi die enkel over een CPU beschikt. Ook wordt een Nvidia Jetson Nano en een Personal Computer gebruikt die zowel een CPU als een GPU bezitten.

Om een representatieve benchmark te bekomen is het nodig om verschillende programma's op een identieke wijze te testen. De te beproeven programma's zullen verdeeld worden over de categorieën regressie en classificatie. De data die uit de benchmark afgeleid wordt bestaat uit de latency en het processor-gebruik van het runnen van elk programma. Deze worden dan herwerkt en genormaliseerd om de resultaten eenvoudiger te analyseren. Hierbij worden ook factoren zoals prijs, kloksnelheid en energieverbruik in rekening gebracht.

Uit de resultaten kan er een duidelijk verschil afgeleid worden tussen de verschillende edge-toestellen. Hierbij komt de Nvidia Jetson Nano het best naar voor. Het toestel behaalt de laagste latency voor alle toegepaste programma's. De resultaten geven ook aan dat de Coral Dev efficiënter met energie omspringt. Voor het uitvoeren van hetzelfde programma zal de Coral Dev minder energie verbruiken.

Trefwoorden: benchmark, edge-devices, latency, Jetson Nano, Coral Dev, Raspberry Pi

Lijst van figuren

Lijst van tabellen

Lijst van symbolen en acroniemen

Symbols

λ_{ex}	Excitation wavelength
λ_{em}	Emission wavelength
$N(\ldots)$	Noise process

NOG TE DOEN OP HET EINDE VAN DE THESIS

Acronyms

LPWAN	Low-Power Wide-Area Network
IOT	Internet-of-Things
ISI	Inter-Symbol Interference
SBC	Single Board Computers
ANN	Artificiële Neurale Netwerken
ML	Machine Learning
NN	Neurale Netwerken
SOC	System On Chip
AI	Artificiële Intelligentie
EVE	embedded-vision-engine
SVM	Support Vector Machines
TPU	Tensor Processing Unit
ASIC	Application Specific Integrated Circuit
TOPS	tera operations per second
TOPW	tera operations per Watt
CPU	Central Processing Unit
RELU	rectified linear unit
DLIB	Deep Learning Inference Benchmarks
DL	Deep Learning
CNN	Convolutional Neural Networks
TF	Tensorflow
TFL	Tensorflow Lite
RNN	Recurrent Neural Network
CSV	Comma Seperated Value
RTRL	Real Time Recurrent Learning
BPTT	Back Propagation Trough Time

PCA	Principal Component Analysis
KNN	K-Nearest Neighbors

Inhoudsopgave

Listings

Hoofdstuk 1

Inleiding

De computerwereld maakt de laatste jaren grote stappen op vlak van Machine Learning[1]. Deze vorderingen werden gedreven door onder meer de nieuwste ontwikkelingen op vlak van computerrekenkracht en de vierde industriële revolutie[2]. Door de veelvuldige toepassingsmogelijkheden werd Machine Learning (ML) een populair en veelbesproken onderwerp. Tegenwoordig kan een bepaalde vorm van Artificiële Intelligentie (AI) in elke sector teruggevonden worden[3]. Van hartritmestoornissen herkennen in de medische sector tot commentaarherkenning in de toeristische stiel.

In deze thesis zal men trachten om een benchmark met AI op te stellen waarmee verschillende Single Board Computers (SBC)s met elkaar vergeleken kunnen worden. Dit instrument moet meer inzicht verschaffen in welke mate machineleertechnieken toepasbaar zijn. Er zal vooral gekeken worden naar performantie-parameters. Hoe groot is de latency die optreedt? Welk verbruik en complexiteit van het netwerk gaat er hier mee gepaard? Ook tussen deze hardware-opties wordt er een afweging gemaakt. Welk toestel is voordeliger in welke situatie? Is een goedkoper toestel tot evenwaardige resultaten in staat? Er zal een oplijsting gemaakt worden van de belangrijkste parameters en met behulp van de benchmark-resultaten zal er een besluit over de SBCs genomen worden.

Hoofdstuk 2

Literatuurstudie

In dit hoofdstuk zal er eerst een afgrenzing gegeven worden waarbinnen de thesis gekozen is. Vervolgens zal een korte introductie gegeven worden tot Artificiële Intelligentie en Machine Learning waarin verschillende vormen en mogelijke toepassingsdomeinen behandeld worden. De verschillende leermodellen worden bekeken en er wordt nagegaan hoe de onderdelen een werkend geheel vormen. Vervolgens wordt er een overzicht van de voor- en nadelen van de verschillende technieken gegeven. Er wordt een best bruikbare methode voor de toepassing in deze thesis gekozen. Bij deze keuze zullen de voor- en nadelen gewikt en gewogen worden. Verder wordt er ook de geschiedenis van Single Board Computers aangehaald. Hoe zijn deze toestellen ontstaan, hoe zijn ze geëvolueerd en in welke staat zijn de hedendaagse SBCs. Verder worden ook verschillende voorbeelden van SBCs besproken die in staat zijn om Machine Learning-technieken toe te passen. Deze zullen onderworpen worden aan een benchmark die in paragraaf 2.6 besproken zal worden.

2.1 Kadering

Een branche in de ML die steeds meer in de schijnwerpers staat, is de logistieke sector[4]. Door de steeds verder doorgedreven automatisatie van bedrijven, wordt er ook in bijvoorbeeld magazijnen geopteerd voor het optimaliseren van onder meer het leveren van de verschillende onderdelen en de veiligheid in het magazijn. Gebruik maken van zelfrijdende vorkheftrucks is een mogelijke optie in het verbeteren van de efficiëntie. Niet alleen in het magazijn maar ook op de weg is er een groeiende belangstelling naar zelfrijdende voertuigen, die ontwikkeld worden door grote bedrijven zoals Tesla en Uber. In beide cases zal er een zekere vorm van positiebepaling nodig zijn. Het is van groot belang dat deze bepaling zo accuraat mogelijk plaats vindt, met niet alleen een juiste locatie, maar ook een resultaat dat op zo kort mogelijke tijdsperiode geproduceerd wordt.

Deze nood aan *low latency* kan het verschil betekenen tussen een voertuig dat beslist dat hij moet vertragen of beslist dat hij veilig kan doorrijden maar toch een botsing veroorzaakt. De berekening van die cruciale locatiebepaling kan zowel in de *cloud*, als in de *edge*[5] gebeuren en gebeurt onder meer met behulp van ML. Hierbij wordt met cloud verwezen naar het verwerken van data op een locatie ver weg van het voertuig zoals serverzalen. Met edge wordt dan weer een locatie dichtbij het voertuig bedoeld. Dit kan zowel op, als vlakbij het voertuig zijn. *Cloud computing* brengt een zekere toegevoegde latency teweeg. Dit is de nodige tijd om via het internet de server te bereiken. Hierdoor zal men eerder kiezen voor *edge computing* vallen. Indien de berekeningen in de edge plaats vinden zal er bijvoorbeeld een kleinere vertraging zijn tussen het moment van vertrekken en het waarnemen door het algoritme dat er vertrokken werd. De afstand tussen de reële locatie, waar het voertuig zich fysisch ook bevindt, en de virtuele locatie, waar de computer het voertuig acht te zijn, is kleiner bij lagere latencies. Dit leidt tot een betere positiebepaling. Deze heeft dan weer tot gevolg dat meer ongevallen vermeden kunnen worden. Welke hardware men gebruikt kan variëren van applicatie tot applicatie. De berekening zelf wordt uitgevoerd met behulp van AI doordat dit verschillende baten heeft. Deze voordelen worden besproken in de volgende paragrafen.

2.2 Artificiële Intelligentie

AI verwijst naar het simuleren van menselijk intellect in machines die geprogrammeerd worden om de menselijke redenering na te bootsen. Het wordt gezien als de studie van *intelligente agents* die zijn omgeving waarnemen en acties kunnen ondernemen om de kans van het bereiken van een bepaald doel te maximaliseren[6]. Er kan een onderscheid gemaakt worden tussen zowel sterke als zwakke AI.

- **Zwakke AI:** Dit is een vorm van AI die zich bezig houdt met onderzoek in gebieden waar handelswijzen mogelijk zijn die tekenen van intelligentie vertonen, maar niet volwaardig intelligent zijn. Hier worden de meeste vorderingen in voortgebracht, zoals handschriftherkenning of zoekalgoritmen.
- **Sterke AI:** Deze vorm van AI houdt zich bezig met onderzoek dat als doel heeft om software te creëren die zelfstandig kan redeneren en problemen aanpakken.

Het gebruiken van AI kan vele voordelen hebben[7][8]. Zo is het mogelijk om data beter en sneller te gebruiken dan de mens kan. Data kan gelezen en verwerkt worden in geautomatiseerde processen zonder de tussenkomst van een persoon en in een fractie van een seconde. Verder zal er ook, indien er meer beschikbare data zijn, een nog nauwkeuriger responsie gegeven worden. AI wordt als zwakke AI veel toegepast om repetitieve taken met relatief lage complexiteit over te nemen. Een belangrijk onderdeel is Machine Learning dat verder uitgelegd zal worden in volgende paragraaf.

2.3 Machine Learning

ML is een onderdeel van AI en is de studie en modellering van de verschillende leerprocessen dat gebruikt kan worden door verschillende computersystemen[9]. Deze systemen zijn hierdoor in staat om specifieke taken te voltooien zonder rechtstreekse instructies of regels mee te krijgen van de operator. Ze steunen in de plaats op onder meer patroonherkenning om de kans op het succesvol uit te voeren van taken te maximaliseren. Hiervoor wordt er een wiskundig model gebouwd dat gebruik maakt van trainingsdata. Deze mathematische modellen en *datahandling* kan op verscheidene manieren gebeuren. In dit hoofdstuk worden eerst een aantal algemene leertechnieken uitgelegd. Vervolgens worden een aantal gebruikelijke modellen besproken. Tot slot wordt een afweging gemaakt over welk model het interessantst is voor onze toepassing.

2.3.1 Leertechnieken

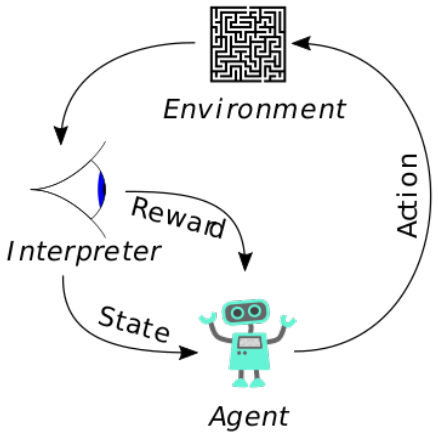
Er zijn verschillende mogelijkheden om een neuraal netwerk te laten leren. De drie belangrijkste methodes om een mathematische functie te verkrijgen worden hieronder opgesomd[3].

Supervised Learning

Deze techniek maakt gebruik van gepaarde datasets van inputobjecten en de te verwachten outputobjecten. Het doel is om een mathematische functie te creëren waarbij de gegenereerde outputs zo nauw mogelijk overeenkomen met de gelabelde outputs uit de datasets. Men optimaliseert deze mathematische functie door iteratief te trainen. De bijgeschaafde functie kan dan ook gebruikt worden voor nieuwe datasets zonder gelabelde output. Hierbij zal hij zelf outputwaardes genereren. De meest toegepaste werkwijzes zijn lineaire regressie, beslissingsbomen en Neurale Netwerken (NN). Een toepassing van Supervised Learning is bijvoorbeeld het detecteren van spam met een trainingset van al gelabelde e-mails.

Unsupervised Learning

Unsupervised learning is een techniek die gebruik maakt van Hebbian Learning om onbekende patronen te herkennen in datasets. De meest gebruikte methode onder Unsupervised Learning zijn is cluster analyse. Hierbij wordt er getracht om groep objecten te identificeren en te verdelen in clusters van gelijkaardige objecten. Deze werkwijze kan op twee voorname manier gebeuren. De eerste en meest gekende is Principle Component Analysis (PCA). PCA maakt gebruik van orthogonale transformaties om een set van mogelijke afhankelijke variabelen om te zetten in een set van lineaire onafhankelijke variabelen. Een tweede werkwijze is met behulp van K-Nearest Neighbors (KNN). Hierbij wordt er een onderscheid gemaakt tussen clusters door gebruik te maken van k nabije punten om een clusters te identificeren. Een belangrijke toepassing van Unsupervised Learning is het clusteren van gelijkaardige documenten op basis van de inhoud van de tekst.



Figuur 2.1: Routine bij Reinforcement Learning.[10]

Reinforcement Learning

Deze leertechniek heeft betrekking tot hoe agents acties moeten ondernemen in een omgeving om een bepaald attribuut te maximaliseren. Het onderscheidt zich van Supervised en Unsupervised Learning door de onafhankelijkheid van gelabelde outputdatasets. De techniek heeft als doel om een evenwicht te vinden tussen exploratie van ongekend gebied en exploitatie van de huidige kennis. In figuur 2.1 kan je een eenvoudige routine vinden van Reinforcement Learning-algoritme. Hierbij maakt een agent een bepaalde actie gebaseerd op de staat waar hij in is. Deze actie heeft in een omgeving een zekere invloed die door een Interpreter beoordeeld wordt en een score toekent. De agent kan deze verandering daarna gebruiken om zichzelf te verbeteren en zijn acties aanpassen.

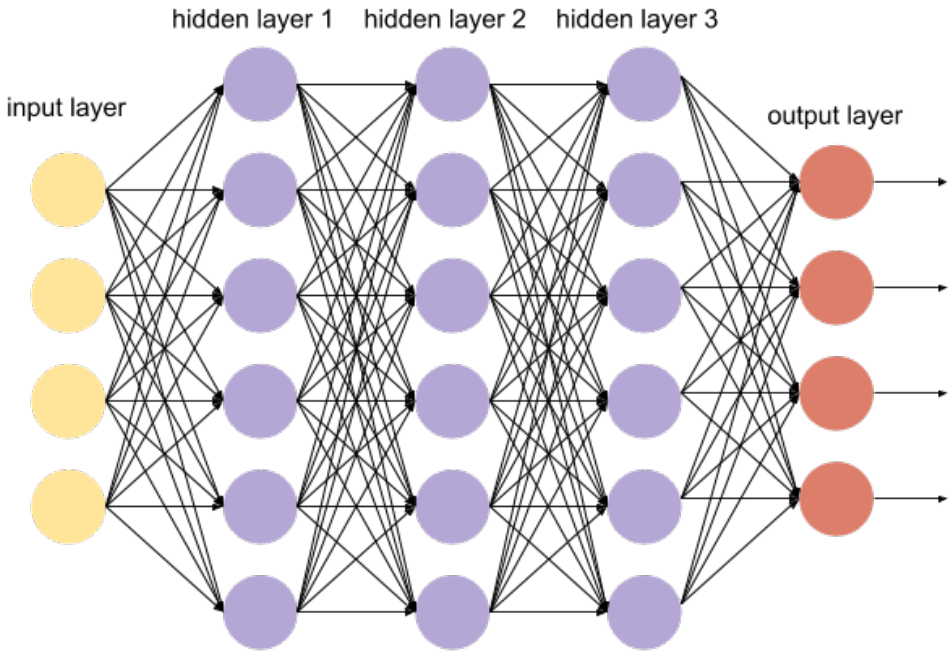
Een van de vele mogelijke toepassingen van Reinforcement Learning is het aanleren van schaken door enkel mee te geven of het algoritme gewonnen of verloren heeft.

2.3.2 Leeralgoritmes en -technieken

Om ML technieken toe te passen moet men gebruik maken van een bepaald wiskundig model dat is getraind op trainingsdata en hierdoor nieuwe data kan verwerken om voorspellingen te maken. Er bestaat een hele waaier aan mogelijke modellen. In de volgende paragrafen worden een aantal opties besproken waarna er de verschillende modellen met elkaar vergeleken worden.

Artificiële Neurale Netwerken

Artificiële Neurale Netwerken (ANN) zijn een term dat gebruikt wordt om algoritmes te omschrijven die in staat zijn om een bepaalde taak te leren, en zichzelf te verbeteren. In de meeste gevallen worden er amper richtlijnen of een omschrijving meegegeven. Het systeem ontdekt zelf hoe deze regels in elkaar zitten[9]. Hierbij blijft de interpretatie van de input en de output wel nog belangrijk. Een bekend voorbeeld is het herkennen van de cijfers 0 tot 9. Hier wordt er niet aan het systeem verteld hoe de vorm van een getal er uit ziet. Het algoritme zal dit gaandeweg ontdekken, met behulp van vele voorbeelden waar het gebruik van kan maken. Met behulp van veel data kan een algoritme zichzelf verfijnen en zo nauwkeuriger bepaalde cijfers herkennen.



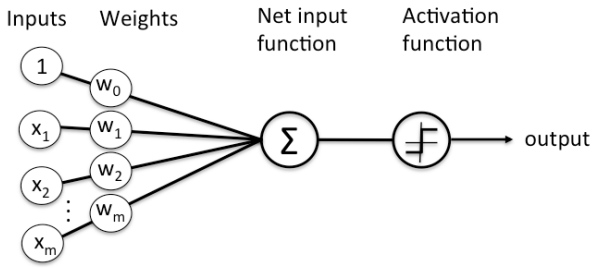
Figuur 2.2: Structuur van een Neuraal Netwerk[11].

Structuur van Neurale Netwerken Een ANN is een verzameling van nodes die met elkaar verbonden zijn zoals neuronen in de hersens van een mens. Hierbij kan elke neuron een signaal doorgeven naar het volgende neuron waar het signaal verwerkt kan worden en weer doorgegeven kan worden. Hetzelfde principe geldt ook bij NN met het verschil dat er meerdere lagen van nodes te onderscheiden zijn.

Lagen Er zijn drie soorten lagen te onderscheiden: een Input Layer, Hidden Layers en een Output Layer. Elke laag is verbonden met de volgende laag door middel van connecties tussen de verschillende nodes. In figuur 2.2 is de algemene vorm van een NN te vinden.

- **Input Layer:** De eerste laag van elke NN is de Input Layer. Deze bestaat uit een aantal inputnodes. Elke inputnode krijgt de ruwe data binnen waar er een operatie op uitgevoerd wordt en vervolgens bepaalde parameters doorgeeft aan de volgende laag. De wijze waarop data geïnterpreteerd worden, vormt een belangrijk vertrekpunt voor het NN.
- **Hidden Layers:** Na de inputlaag komen een aantal Hidden Layers. Het aantal Hidden Layers en de hoeveelheid nodes binnen één Hidden Layer kan variëren van applicatie tot applicatie en is sterk gerelateerd aan de complexiteit van de toepassing.
- **Output Layer:** Na de Hidden Layers is de laatste laag de Output Layer. Hier worden de laatste operaties uitgevoerd en worden de eindwaarden verkregen waar het resultaat uit afgeleid kan worden.

Nodes Lagen zijn opgebouwd uit meerdere nodes. Elke node krijgt een bepaald aantal inputs, verwerkt deze en geeft een bepaald aantal outputs. Deze inputs en outputs worden van node



Figuur 2.3: Algemene structuur van een node[12].

naar node doorgegeven via verbindingen. Elke node heeft met elke node in de volgende laag een connectie. Elke verbinding draagt een bepaald gewicht. Via dit gewicht kan men de invloed van de huidige node versterken of verzwakken in de volgende node.

Activatie functie De mathematische functie die een node gebruikt voor het verwerken van inputs naar outputs heet de activatie functie. In figuur 2.3 wordt de algemene vorm van een neuron besproken. Deze neuron heeft $m+1$ inputs $(x_0\ t.e.m.\ x_m)$ en bijhorende gewichten $(w_0\ t.e.m.\ w_m)$. Gebruikelijk wordt $x_0 = +1$ genomen. Hierdoor blijven er maar m echte inputs over waardoor er voor een bepaalde output de functie 2.1 opgesteld kan worden. Hierbij is ϕ een van de mogelijke transferfuncties die verder besproken zal worden.

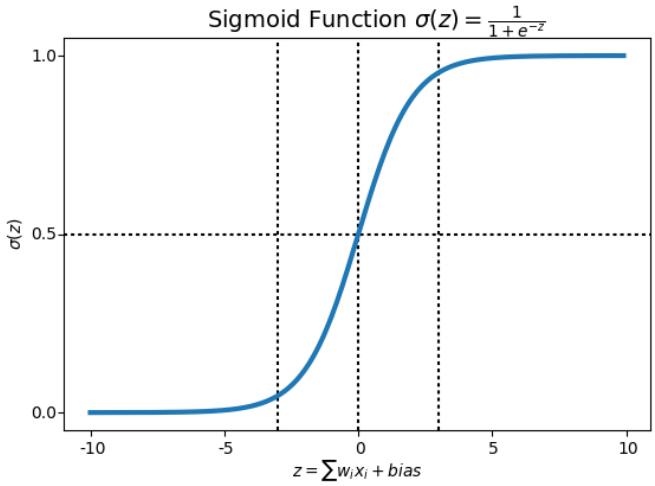
$$y = \phi\left(\sum_{j=0}^{m} w_j x_j\right) \tag{2.1}$$

Types activatiefuncties De transfer functie of activatiefunctie is een belangrijk onderdeel van een laag in een ANN. De activatiefunctie transformeert inputs uit een vorige laag en transformeert deze naar een output. Deze output zal voor een volgende laag weer als input gebruikt worden. Door gebruik te maken van de juiste activatiefuncties kunnen er niet-lineaire eigenschappen aan het netwerk toegevoegd worden. Hieronder zullen enkele transferfuncties besproken worden.

- **Lineare Combinaties:** In dit geval is de output niets minder dan de gewogen som vermenigvuldigd met een constante waarbij zoals in formule 2.2 waar er nog een tweede constante wordt opgeteld.
- **Stapfunctie:** Hier wordt er gekeken naar de verkregen waarde van de gewogen som u van $m+1$ inputs. Bedraagt deze waarde minder dan een bepaalde drempel θ, dan wordt de output gelijkgesteld aan nul, bij een hogere waarde dan weer aan 1. Dit is te zien in formule 2.3. Dit type wordt vooral gebruikt om binaire inputs te verzorgen bij de volgende laag.

$$u = \sum_{j=0}^{m} w_{kj} x_j \tag{2.2}$$

$$y = \begin{cases} 1 & \text{als } u \geq \theta \\ 0 & \text{als } u < \theta \end{cases} \tag{2.3}$$



Figuur 2.4: De Sigmoid activatiefunctie[15].

$$y = \frac{1}{1+e^{-u}} \tag{2.4}$$

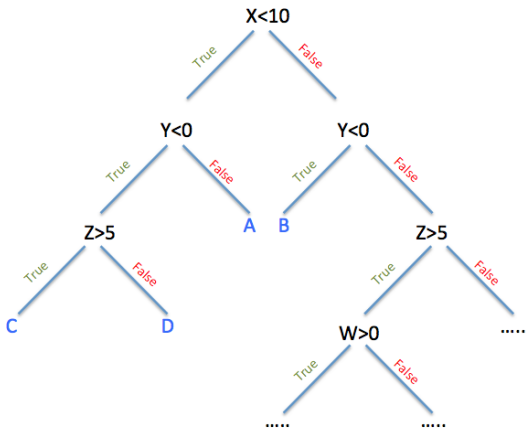
- **Sigmoid:** De Sigmoid functie[13] is een mathematische functie zoals te zien is in figuur 2.4. Het heeft de karakteristieke 'S'-vorm zoals te zien is in figuur 2.4. Door deze activatiefunctie worden inputs omgezet in een waarde tussen 0 en 1 (of -1 en 1, afhankelijk van de conventie).
- **Rectifier:** De rectifier als activatiefunctie[14] is een functie die enkel het positieve deel van zijn argument doorlaat. In vergelijking 2.5 vind je de functie weer waar x de input is van de neuron. Deze is een vector aan waarden die zowel positief als negatief kunnen zijn. Deze functie is ook gekend onder de naam *rectified linear unit (ReLU)*.

$$f(x) = x^+ = max(0,x) \tag{2.5}$$

Beslissingsboom

Het gebruik van een beslissingsboom is een leermethode die met regelmaat terugkomt in de statistiek als een voorspellend model. Men maakt gebruik van observaties rond een bepaalde uitspraak. Men kwantificeert deze observaties zodat deze leiden naar een variabele outputwaarde. Indien deze waarde valt onder te verdelen in discrete klassen spreekt men over een *classificatie boom*. Neemt de gezochte variabele eerder een continue vorm aan, dan maakt men gebruik van *regressie bomen*.

Structuur van een beslissingsboom Net zoals bij de NN bestaat een beslissingsboom uit verschillende lagen en nodes. Zoals te zien is in figuur 2.5 wordt er in elke laag een onderscheid gemaakt op basis van een statement of parameter. Deze parameter kan een enkele inputwaarde zijn of een lineaire combinatie van meerdere inputwaarden.



Figuur 2.5: Algemene structuur van een beslissingsboom[16].

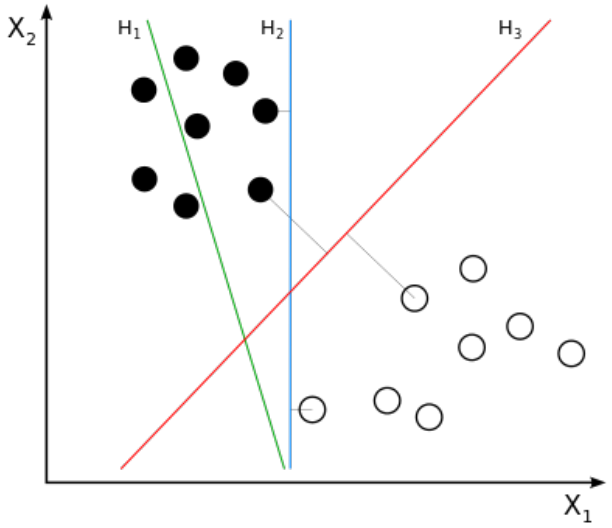
Bij de nodes kan er onderscheid gemaakt worden tussen een gewone node en een eindnode. Bij elke gewone node wordt een bepaald statement geverifieerd en wordt er naar een node overgegaan in de volgende laag op basis van dit statement. Bij een eindnode is het niet meer mogelijk om door te gaan naar een volgende laag, maar wordt er een outputwaarde gegeven.

Support Vector Machines

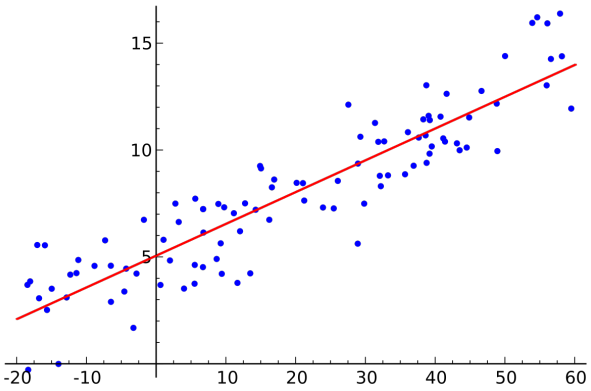
Support Vector Machines (SVMs) of ook wel gekend als support vector networks is een model dat veelvuldig gebruikt wordt in classificatie en regressie-analyse. Een belangrijke toepassing van SVMs is het verdelen van objecten in twee verschillende klassen op basis van een aantal kenmerken. Het is dus onder meer een binaire classificeerder. Om aan de hand van kenmerken een onderverdeling te maken moeten deze kenmerken eerst omgezet worden in een vectorruimte. In de trainingsfase wordt er getracht een zo optimaal mogelijke scheiding tussen beide klassen te vinden. Deze optimale scheiding wordt ook wel een *hypervlak* genoemd en ligt op een zo groot mogelijke afstand tussen de dichtstbijgelegen objecten van beide klasses of support vectors. in figuur 2.6 kan u een tweedimensionaal voorbeeld vinden. Hierin is scheidingslijn H1 geen acceptabele scheiding omdat er objecten van de zwarte klasse fout geclassificeerd worden. H2 is acceptabel maar is nog niet optimaal aangezien er weinig foutmarge is voor een nieuw object. H3 is het hypervlak omdat de foutenmarge tussen de twee klassen zo groot mogelijk is. Deze methode is niet alleen bruikbaar in toepassingen met een lineaire scheiding. Ook in niet-lineaire gevallen kan men een transformatie uitvoeren om toch een lineaire scheiding te bekomen. Deze hervorming wordt ook wel de *kernel trick*[17] genoemd.

Regressie Analyse

Regressie Analyse is een techniek uit de statistiek[19], die gebruikt wordt om gegevens te analyseren met een specifiek verband. Er bestaat vaak een relatie tussen een afhankelijke variabele en één (of meerdere) onafhankelijke variabelen. De meest voorkomende vorm van regressie analyse is de lineaire regressie, waar men op zoek gaat naar de functie die het dichtst aanleunt bij de data. De functie moet wel vervullen aan specifieke criteria, zo moet de functie bijvoorbeeld een bepaalde



Figuur 2.6: Tweedimensionale Support Vector Machine[18].



Figuur 2.7: Voorbeeld van Lineaire Regressie[20].

orde hebben. Regressie analyse wordt vooral gebruikt voor het voorspellen van nieuwe data of gebeurtenissen. In figuur 2.7 kan je een voorbeeld van lineaire regressie vinden.

Bayesian Netwerken

Bayesian Netwerken, ook wel probabilistische netwerken genoemd, zijn structuren waarin data op probabilistische wijze geanalyseerd kunnen worden. Dit wil zeggen dat men als output niet enkel de outputs maar ook de onzekerheid hierop krijgt. Men maakt gebruik van gerichte grafen. Hierin bestaan de knopen uit variabelen en de arcs beschrijven de conditionele afhankelijkheden tussen de verschillende knopen. Bayesian Netwerken worden vooral gebruikt om te analyseren wat de bepalende oorzaak is voor een zekere gebeurtenis.

Genetische Algoritmes

Genetische Algoritmes zijn een heuristiek geïnspireerd op het principe van natuurlijke selectie en zijn een klasse binnen evolutionaire algoritmes. Dit type algoritme kan gebruikt worden om

oplossingen te vinden in optimalisatie- en zoekproblemen. Door te steunen op biologische principes zoals mutatie, selectie en kruisbestuiving worden er nieuwe *chromosomen* gegenereerd die mogelijks een betere oplossing geven voor een bepaald probleem.

2.3.3 Keuze voor een Machine Learning-methode

Om de keuze voor een bepaalde ML-techniek te verantwoorden, zal dit hoofdstuk de voor- en nadelen behandelen van de voornaamste technieken die via regressie of classificatie toegepast kunnen worden[21].

Regressie

- **Neurale Netwerken:**

  Voordelen: Neurale netwerken zijn de meest gebruikte toepassing in verschillende domeinen. NN kunnen uitstekend omgaan met onder meer beeld-, audio- en tekstdata en deze verwerken. Verder kan de architectuur ook nog gemakkelijk aangepast worden aan de toepassing door te variëren in het aantal lagen of nodes. Het gebruik van de hidden layers vermindert ook het hanteren van feature engineering.

  Nadelen: Neurale netwerken zijn minder bruikbaar voor *general-purpose* algoritmes door de grote hoeveelheid data die er voor nodig zijn. In dat geval is het beter om voor beslissingsbomen te kiezen. Bovendien vragen ze veel computationeel vermogen voor het trainen van het netwerk en vragen veel expertise voor kleine aanpassingen zoals aan de architectuur of hyperparameters.

- **Regression Trees:**

  Voordelen: Beslissingsbomen met nadruk op regressie zijn in staat om niet-lineaire relaties te leren en zijn robuust voor uitschieters in de te verwerken dataset.

  Nadelen: Regression trees zijn vatbaar voor overfitting indien er te veel gebruik gemaakt wordt van branches. Bij bomen is het ook mogelijk om vertakkingen te blijven maken tot het een exacte kopie voorstelt van de trainingsdata.

- **Lineaire Regressie:**

  Voordelen: Dit is een eenvoudige methode om zowel te begrijpen als uit te leggen. Daarnaast kan er een eenvoudige bescherming tegen overfitting geïmplementeerd worden.

  Nadelen: Niet-lineaire relaties zijn een zwak punt voor lineaire regressie. Het is moeilijk om een correcte fitting te vinden voor een gegeven ingewikkelde relatie. Bovendien is het onvoldoende flexibel om complexe patronen op te vangen.

Classificatie

- **Neurale Netwerken:**

  Voordelen: NN blijven uitstekend presteren bij het classificeren van audio-, tekst- en beeldherkenning.

  Nadelen: Er is nood aan grote hoeveelheden data om het model te trainen en minder geschikt als general-purpose algoritme.

- **Classification Trees:**

  Voordelen: Verrichten zeer goed werk in praktijk. Ze zijn robuust voor uitschieters, schaalbaar voor meerdere klassen en kunnen niet-lineaire grenzen op natuurlijke wijze modelleren dankzij de hiërarchische structuur.

  Nadelen: Classification trees zijn vatbaar voor overfitting indien er te veel gebruik gemaakt wordt van branches. Bomen hebben vaak de neiging om branches aan te maken tot het een exacte kopie voorstelt van de trainingsdata.

- **Support Vector Machines:**

  Voordelen: SVMs zijn in staat om niet-lineaire beslissingsgrenzen te modelleren en hebben een sterke robuustheid tegen overfitting, vooral in hogere dimensionale vectorruimtes.

  Nadelen: SVMs zijn heel erg geheugen intensief. Ze vragen ook meer expertise in het afstemmen door het grote aanbod in mogelijke kernels. SVMs hebben de eigenschap om minder effectief te zijn bij het schalen naar grotere datasets.

- **Geregulariseerde Regressie:**

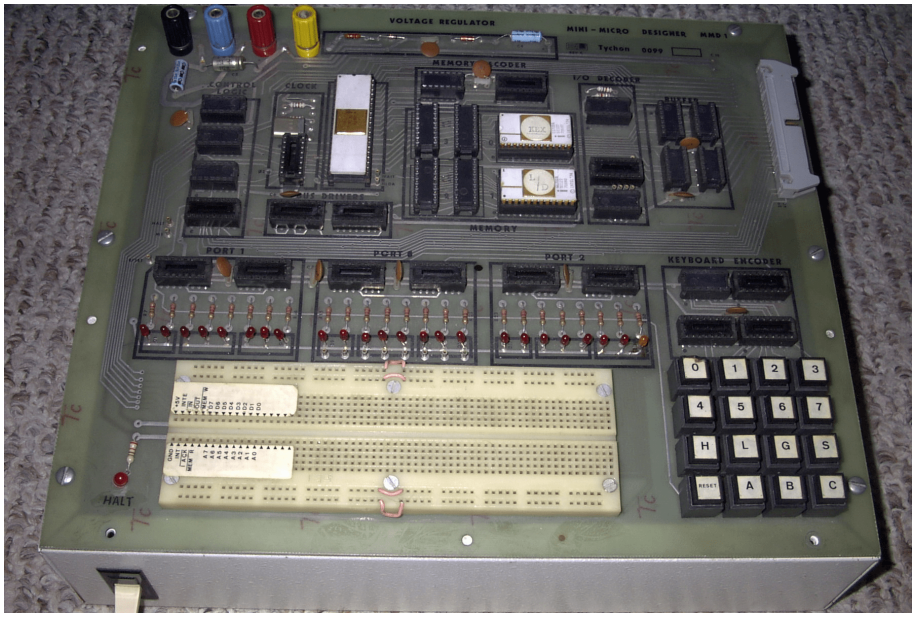
  Voordelen: Outputs hebben een gemakkelijk leesbare probabilistische interpretatie. Ook kan er bescherming tegen overfitting geïmplementeerd worden en kunnen modellen eenvoudig geüpdatet worden.

  Nadelen: Niet-lineaire relaties zijn een zwak punt voor lineaire regressie. Het is moeilijk om een correcte fitting te vinden voor een gegeven ingewikkelde relatie. Bovendien is het onvoldoende flexibel om complexe patronen op te vangen.

Besluit

Na een afweging gedaan te hebben van de voor- en nadelen van de voornaamste kandidaten bij zowel regressie als classificatie toepassingen zal in het kader van deze thesis voor NN gekozen worden. Er wordt vooral gesteund op het feit dat NN uitstekend werk levert in beide klassen in het analyseren van data. Bovendien zijn de voornaamste nadelen minder van toepassing in het kader waarin NN toegepast zal worden. Het is vooral van belang hoe de netwerken in de executiefase presteren. Het trainen van de verschillende netwerken met grote hoeveelheden data kan op aparte systemen gebeuren. De trainingsfase is dus minder van belang voor het doel van deze thesis. Bovendien is het niet nodig om een breed general-purpose netwerk te voorzien.

Een mogelijk alternatief voor NN is het gebruiken van Regressie Analyse. De bepalende factor hiervoor is dat het karakter van de verscheidene applicaties vaak uit lineaire relaties bestaat.



Figuur 2.8: Eerste SBC: MMD-1[22].

2.4 Evolutie Single Board Computers

Een SBC is een volledige computer gemaakt op 1 enkele printplaat. Het bevat onderdelen zoals een microprocessor, geheugen, inputs en outputs. De eerste SBC werd ontwikkeld als een voorstel-hulpmiddel bij educatieve doelstellingen of het gebruik als een embedded computer controller. Tegenwoordig zijn ook vele (draagbare) computers geïntegreerd op één printplaat. Het grote verschil met (draagbare) computers is dat er geen nood is aan expansion slots zoals bijvoorbeeld voor RAM-geheugen of een Graphics Processing Unit (GPU).

2.4.1 Geschiedenis

De eerste echte SBC was de zogenaamde ”dyna-micro” uit figuur 2.8 die later de naam ”MMD-1”(Mini-Micro Designer 1) kreeg[22]. Dit toestel werd uitgegeven in 1976 en werd populair doordat het werd gepresenteerd in het destijds ’BugBook’. Een andere vroege SBC was de KIM-1 (Keyboard Input Monitor 1) uit hetzelfde jaar. Beide machines werden voor ingenieurs geproduceerd en ontworpen maar vonden een breed publiek onder de hobbyisten waar het heel populair werd. Later kwamen nog andere namen zoals de Ferguson Big Board en de Nascom.

Naarmate de markt voor desktops en PC’s groeide, nam de belangstelling voor SBC in computers meer en meer af. De focus van de markt werd verlegd naar een moederbord met de belangrijkste componenten en dochterborden voor periferiecomponenten zoals seriële poorten. De voornaamste reden hiervoor was dat de componenten groot waren. Alle onderdelen op dezelfde printplaat zou zorgen voor een onpraktisch ontwerp met grote afmetingen. Deze beweging was echter tijdelijk en naarmate de vorderende technologie kleinere componenten kon leveren, werden onderdelen terug naar het mainframe verschoven. Tegenwoordig kunnen de meeste moederborden terug als SBC beschouwd worden.

In het jaar 2004 werd er in Italië een nieuwe microcontroller uitgebracht onder de naam ”Arduino”. Dit ontwerp had, naast het voordeel van compact en goedkoop te zijn, ook nog eenvoudigheid

mee. Door de eenvoud werd het Arduino-platform snel populair onder techneuten van alle soorten. Twee jaar later bracht de Universiteit van Cambridge een nieuwe goedkope SBC uit. De bekende Raspberry Pi werd gelanceerd voor de prijs van $35. Het hoofddoel van dit project was een nieuw leermiddel om te programmeren maar werd door het grote aantal applicaties ook zeer populair.

De laatste jaren kende een grote explosie aan nieuwe SBCs. Een hele reeks nieuwe namen verschenen. Banana Pi, Beaglebone, Intel Galileo, Google Coral Dev en Asus Tinker Board zijn maar enkele van de vele voorbeelden. Deze toestellen hebben vaak een processor gebaseerd op de x86- of ARM-series en maken gebruik van een Linux besturingssysteem zoals Debian.

2.5 Assortiment aan 'off the shelf' toestellen

In deze sectie wordt er een kort overzicht gegeven van de te gebruiken SBCs binnen deze thesis. Er werd gekozen om met vijf verschillende toestellen te werken die in de edge toegepast kunnen worden. Daarnaast wordt er er ook met een personal Computer gewerkt. De gebruikte devices verschillen in verscheidene aspecten. Zo zijn er zowel goedkope als kostelijkere apparaten, populaire boards als minder gekende SBCs. Er zijn toestellen met hele geavanceerde processoren die specifiek voor ML zijn ontworpen, maar ook processoren die ontwikkeld zijn voor meer algemenere toepassingen. De specificaties van de verscheidene toestellen worden ook meegegeven. In tabel 2.1 kan er een samenvatting van de belangrijkste specificaties gevonden worden.

2.5.1 Beaglebone AI

BeagleBone Ai (BB AI) is een SBC dat verder bouwt op de succesvolle BeagleBoard-series[23]. Het is een open source project met een op Linux gebaseerd aanpak. De BB AI probeert het gat tussen kleinere SBC en krachtigere industriële computers te overbruggen. Met behulp van de krachtige Texas Instruments AM5729 CPU kunnen ontwikkelaars de krachtige System On Chip (SoC) gebruiken om een hele brede waaier aan toepassingen te verwezenlijken. De BB AI maakt het toegankelijker om het AI-terrein te ontdekken en te verkennen. Door gebruik te maken van onder andere embedded-vision-engine (EVE) cores die steunen op een geoptimaliseerde TIDL machine learning OpenCL API, kan je terecht in alledaagse automatisatie in industriële, commerciële en thuisapplicaties.

Specificaties

- **GPU:** Niet van toepassing
- **CPU:** Texas Instruments AM5729
- **Memory:** 16 GB on-board eMMC flash
- **Storage:** 1GB RAM + micro SD-slot
- **Power:** 5 Watt
- **Prijs:** $139.37

2.5.2 Coral Dev Board

De Coral Dev Board is een development board gemaakt door het Amerikaanse technologiebedrijf Google[24]. Het board is ontworpen om het ontwikkelen van on-device ML producten te vergemakkelijken. Hiervoor heeft het een aantal belangrijke voordelen gekregen door zijn designers. Het is vooral de aangepaste Tensor Processing Unit (TPU) AI chip die hier opvalt. De TPU is een Application Specific Integrated Circuit (ASIC) speciaal ontworpen voor NN ML, en is in staat om video in hoge resolutie te analyseren aan 30 frames per second. Deze System-on-Module (SoM) is geoptimaliseerd om Tensorflow Lite te kunnen draaien aan meerdere tera operations per second (TOPS).

Specificaties

- **GPU:** Integrated GC7000 Lite Graphics
- **CPU:** NXP i.MX 8M SOC (quad Cortex-A53, Cortex-M4F) + coprocessor Google Edge TPU
- **Memory:** 8 GB on-board eMMC flash
- **Storage:** 1GB RAM LPDDR4 + micro SD-slot
- **Power:** 0.5 watts for each TOPS - 2 Watt
- **Prijs:** $149.99

2.5.3 Nvidia Jetson Nano

De Jetson Nano is een populair bord uit de Jetson Series van Nvidia[25]. Het is een kleine maar krachtige computer ontwikkeld voor embedded applicaties en low-power AI-Internet-Of-Things (IOT). Deze SBC wordt ondersteund door meerdere bibliotheken in sectoren zoals deep learning, computer vision, beeld en multimedia. De hardware bevat zowel een GPU als een Central Processing Unit (CPU). De GPU bestaat uit een krachtige Maxwell architectuur die beeld kan decoderen aan 500 MP/sec. De CPU is van het type Cortex-A57 met 4 kernen.

Specificaties

- **GPU:** 128-core Maxwell met 128 CUDA-cores
- **CPU:** Quad-core ARM A57 @ 1.43 GHz
- **Memory:** 4 GB 64-bit LPDDR4 25.6 GB/s
- **Storage:** micro SD-slot
- **Power:** 5 - 10 W
- **Prijs:** $99

2.5.4 Nvidia Jetson TX2

De TX2, uit de zelfde Jetsonserie, is de high end versie van de hiervoor besproken Nano[26]. Het is het snelste en meest power-efficiëntst van de embedded AI toestellen die gebruikt zal worden in deze thesis. De TX2 verbruikt een 7.5 Watt en brengt het zware AI-rekenwerk naar de edge. Zijn bekwame GPU met 256 CUDA kernen en een duo CPU zijn in staat om de meest geavanceerde Machine Leertechnieken uit te voeren. De grote geheugenvoorzieningen zorgen bovendien dat de datasetgrootte geen beperkende factor meer kan spelen. Verder wordt er nog gezorgd voor een grote ondersteuning via een grote variatie aan hardware interfaces. Hierdoor wordt het integreren van producten aanzienlijk makkelijker.

Specificaties

- **GPU:** 256-core NVIDIA Pascal GPU architecture with 256 NVIDIA CUDA cores
- **CPU:** Dual-Core NVIDIA Denver 2 64-Bit & CPU Quad-Core ARM Cortex-A57 MPCore
- **Memory:** 8GB 128-bit LPDDR4 Memory 1866 MHx - 59.7 GB/s
- **Storage:** 32GB eMMC 5.1
- **Power:** 7,5 - 15 W
- **Prijs:** $399

2.5.5 Raspberry Pi

De laatste SBC die hier besproken wordt, is de Raspberry Pi 4[27]. Dit is een heel goedkoop en eenvoudig device. Het komt uit de heel gekende Raspberry-reeks zoals al besproken in paragraaf 2.4.1. In deze thesis wordt gebruik gemaakt van het model 3 B. De specificaties kunnen hieronder gevonden worden. Het heeft een behoorlijke Cortex Quad core-CPU met degelijk geheugen die uitgebreid wordt door een SD-kaart.

Specificaties

- **CPU:** Broadcom BCM2711, Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz
- **Memory:** 1GB, 2GB or 4GB LPDDR4-3200 SDRAM
- **Storage:** micro SD-kaart
- **Power:** 2,8 - 5,2 W
- **Prijs:** $35

Toestel	GPU	CPU	Power [W]	Price [$]
Beaglebone Ai	n.v.t.	TI AM5729	5	139,37
Coral Dev Board	GC7000 Lite Graphics	NXP i.MX 8M + Edge TPU	2	149,99
Jetson Nano	128-core Maxwell	Quad-core ARM A57	5 -10	99
Jetson TX2	256-core Pascal	Denver 2 & Cortex-A57	7,5-15	399
Raspberri pi	n.v.t.	BCM2711, Cortex-A72	2,8 - 5,2 W	35
PC	GeForce GTX 1050 Ti	Intel Core i5-7300HQ	10 - 100 W	981

Tabel 2.1: Specificaties van gebruikte toestellen.

2.5.6 Personal Computer: Lenovo Legion Y520

De gebruikte personal computer is van het technologiemerk Lenovo. De Legion-serie biedt een breed gamma gamingcomputers aan het publiek. Het gebruikte Y520-model[28] is samengesteld uit een aantal krachtige onderdelen. Zo bevat het model een sterke GPU, de NVIDIA GTX 1050, en CPU, de Intel Core i5-7300HQ, zorgen in combinatie met een hoge kloksnelheid voor een hoge performantie. De kloksnelheid bedraagt in idle-toestand 2500 MHz en onder load kan dit door gebruik te maken van *dynamic frequency changing* opgedreven worden tot 3300 MHz. De grote hoeveelheden RAM-geheugen en opslagruimte staan toe om meer eisende programma's te runnen op het toestel. Deze grote performantie vraagt wel meer vermogen en een hogere prijs dan de tot nu toe vermelde toestellen. Bij het vermelde vermogen zitten randapparatuur zoals scherm en toetsenbord ook in verwerkt.

Specificaties

- **GPU:** NVIDIA GeForce GTX 1050 Ti Mobile
- **CPU:** Intel Core i5-7300HQ
- **Memory:** 8GB DDR4-2400
- **Storage:** 128 GB SSD, 1 TB HDD
- **Power:** 10 - 100 W
- **Prijs:** $981

2.6 Benchmarking van Machine Learning algoritmes

Om betekenisvolle resultaten te verkrijgen is het nodig om een goed bruikbare en representatieve benchmark op te stellen. Een benchmark is een onderzoek waarbij de prestaties van programma's met elkaar vergeleken worden. Dit komt tot stand als elk programma op identieke wijze wordt onderzocht. Door de prestaties van programma's met elkaar te vergelijken is het mogelijk de performantie van de verscheidene SBC in kaart te brengen. Hoe de benchmark exact in elkaar zit is gebonden aan de kwaliteitscriteria die onderzocht worden. Om ervoor te zorgen dat de benchmark onafhankelijk is van zowel het specifiek veld als toepassing, is het nodig dat er aan een aantal karakteristieken wordt voldaan[29].

- **Vergelijkbaarheid:** Benchmarks moeten zodanig opgesteld zijn, dat het evident is wat ze vergelijken en de conclusie ondubbelzinnig is.
- **Herhaalbaarheid:** Bij het herhalen van de test onder gelijkaardige omstandigheden moeten gelijkaardige resultaten gehaald worden.
- **Goed gedefinieerde methodologie:** De werkwijze, methode en aannames moeten voldoende gedocumenteerd en gestaafd worden.
- **Configureerbaar:** Benchmarks moeten beschikken over parameters die aangepast kunnen worden naar het specifieke probleem dat wordt behandeld.

2.6.1 Bestaande benchmarks

De ontwikkelaars van de Jetson Nano vermelden op de Nvidia-website[30] verschillende Deep Learning Inference Benchmarks (DLIB) waarbij de auteur verscheidene op voorhand getrainde Deep Learning (DL) modellen toepassen op het Nano bord. Deze modellen zijn gebaseerd op een brede waaier aan populaire ML frameworks zoals Tensorflow, Caffe, PyTorch en Keras. Bovendien zijn de applicaties ook gespreid over meerdere toepassingen zoals beeldherkenning, objectdetectie, positiebepaling en anderen.

Ook voor de Jetson TX2 bestaat er een benchmark zoals voorgesteld in [31]. Het gaat over een general-purpose benchmark die een aantal parameters controleert. Het gaat over variabelen zoals Frames per Second (FPS), *inference time*, GPU temperatuur en geheugen verbruik. Met inference time wordt de tijd bedoeld om een berekening te doen in GPU en CPU m.a.w. de latency veroorzaakt door de SBC. Deze gegevens worden uit de data gehaald door middel van verschillende DL modellen toe te passen op Convolutional Neural Networks (CNN). Deze modellen passen ze toe op twee verschillende datasets: Microsoft COCO dataset voor objectdetectie, een foto bank met een grote verscheidenheid aan categorieën, en de KITTI Stereo Vision databank. De KITTI-databank bestaat uit 400 foto's specifiek bedoeld als benchmark fotoset.

Er bestaan ook meer algemenere benchmarks om vergelijkingen tussen computersystemen te maken. Zo bestaat er ook de LINPACK benchmarks[32]. Dit is een programma waar de snelheid gemeten kan worden waarmee een computer in staat is om een n bij n matrix van lineaire vergelijkingen op te lossen. Ondanks dat het een veelgebruikte benchmark is, zijn er wel nog bedenkingen over de werkwijze. Zo bestaat de kerntaak maar uit een enkele computationele taak die onmogelijk de algemene performantie van een systeem kan weergeven. Desondanks geeft de LINPACK benchmark een goed karakteristiek beeld van computersystemen. De meest bekende

toepassing waar LINPACK in toegepast wordt is de ranking van de beste 500 supercomputers ter wereld[33]

Hoofdstuk 3

Data verwerving

In dit hoofdstuk wordt de verwerving van de data toegelicht. Er worden eerst enkele belangrijke softwaretools besproken, die gedurende de thesis gehanteerd zullen worden. Dan worden een aantal specificaties van de gebruikte toestellen besproken. Vervolgens worden de gehanteerde programma's doorgenomen op vlak van toegepaste data en type NN. Het hoofdprogramma wordt in de Python-programmeertaal geschreven. Deze taal werd gekozen door de veelvuldige toepassingsmogelijkheden binnen ML. Tot slot wordt de dataverwerving van de benchmark zelf geïllustreerd en besproken hoe de data opgeslagen wordt.

3.1 Verkennen van software

In deze thesis worden twee belangrijke libraries gebruikt om ML toe te passen op applicaties: pyrenn[1] en TensorFlow (TF)[2]. Beiden zijn een toolbox die toelaten om op heel eenvoudige manieren NN-modellen op te stellen, deze te trainen en te laten uitvoeren.
Pyrenn is de eerste bibliotheek waar gebruik van gemaakt wordt in onderafdeling 3.3.1. Het wordt toegepast op regressie-applicaties en maakt gebruik van het Levenberg-Marquardt algoritm voor het trainen van NN. Bij TF kan dit algoritme gekozen worden uit meerdere opties. De pyrenn-toolbox heeft 2 *dependencies* of afhankelijkheden in Python namelijk: pandas en numpy packages.

De tweede gebruikte bibliotheek is TF. Dit is een gratis open-source software library dat gebruikt wordt om ML-toepassingen uit te voeren, met voornamelijk NN onder de applicaties en Python als gebruikte programmeertaal. Binnen de TF-bibliotheek maken we gebruik van Keras. Dit is net als TF een open-source NN-bibliotheek in Python, maar is niet enkel beperkt tot TF. Ook in andere bibliotheek-omgevingen kan Keras teruggevonden worden. Onder meer in Microsoft Cognitive Toolkit, R, Theano en PlaidML kan dit teruggevonden worden[34]. In deze thesis is er voor gekozen om TF te gebruiken voor de classificatietoepassingen. Bij het opstellen van een NN wordt er gebruik gemaakt van de Keras-bibliotheek. Keras wordt gebruikt i.p.v. TF direct aan te spreken doordat Keras meer gebruiksvriendelijk is.

[1] Meer info over pyrenn is te vinden op https://pyrenn.readthedocs.io/en/latest/index.html
[2] Meer info over TensorFlow is te vinden op https://www.tensorflow.org/

3.2 Verkennen edge-devices

In deze thesis wordt er gebruik gemaakt van drie verschillende edge-devices en een Personal Computer. De drie edge-toestellen zijn: Google Coral Dev Board, Nvidia Jetson Nano en de Raspberry Pi 3. Het zijn alle drie capabele toestellen die heel veelzijdig zijn op vlak van programma's die ze kunnen uitvoeren en randapparatuur dat kan aangesloten worden. In deze sectie worden verschillende eigenschappen van deze toestellen besproken.
De Coral Dev Board is een development board dat ML kan toepassen op de on-board Edge TPU-coprocessor. Om modellen op dit toestel te runnen is er wel nood aan TensorFlow Lite (TFLite)-compatibele modellen. In paragraaf 3.3.3 wordt er uitgelegd hoe een model naar TFLite omgezet kan worden.
Alle devices, die onderworpen worden aan de benchmark, maken gebruik van *dynamic frequency scaling*. Dit is het dynamisch veranderen van de frequentie naargelang de processor veel instructies te verwerken krijgt of niet. Op momenten dat de processor zich in een idle toestand bevindt, kan het gebruik maken van een lagere frequentie om minder vermogen te gebruiken. Door gebruik te maken van het commando $lscpu \mid grep\ MHz$ in een linux-terminal, is het mogelijk om de kloksnelheid weer te geven van het toestel in kwestie. Voor en tijdens het runnen van de benchmark werd dit toegepast om in kaart te brengen welke kloksnelheid werd toegepast op het moment van de benchmark. De Coral Dev gebruikt tijdens de benchmark de 1500 MHz frequentie, in idle toestand is dit 500 MHz. De Nano gebruikt een gelijkaardige kloksnelheid tijdens de benchmark: 1479 MHz. Voor de benchmark bedroeg deze frequentie 102 MHz. De Pi gebruikt dan weer 600 MHz in ruststand en 1200 MHz gedurende de benchmark. De gebruikte Personal Computer benut een klokfrequentie van 2500 MHz in rust en 3250 MHz tijdens de benchmark.
Het verbruik van energie is een belangrijke parameter in het verkrijgen van inzicht in de resultaten. Het verbruikte vermogen voor elk toestel wordt in kaart gebracht. Er kan hiervoor de voedingsvoorziening uit datasheets beschouwd worden. Deze waarden houden echter ook vermogen voor randapparatuur zoals bijvoorbeeld een camera in. In deze thesis zal men vermogenswaarden gebruiken die representatiever zijn. Voor de Coral Dev betekent dit dat het board een vermogen van 2,65 Watt verbruikt. 2 Watt komt van de TPU die 4 TOPS uitvoert aan 2 tera operations per Watt (TOPW). De resterende 0,65 W wordt door de on-board ventilator gebruikt. Voor de Nano kan wel de datasheet voedingswaarde gebruikt worden. Dit komt neer op een verbruik van 10 Watt. De datasheet-waarde heeft hier wel al de randapparatuur in rekening gebracht. Indien er wel randapparatuur aangebracht wordt zal er op een andere manier voeding aan het board geleverd moeten worden. De Pi 3 verbruikt een 3,7 W bij het uitvoeren van programma's zonder randapparatuur. Tot slot verbruikt de Personal Computer 79.9 W bij actief gebruik. Deze waarde is inclusief peripherals zoals scherm, Wi-Fi, muis en toetsenbord aangezien deze niet los te koppelen zijn van de Personal Computer.
Een laatste belangrijke parameter is de kostprijs. Deze werd voor de verschillende toestellen al aangehaald in hoofdstuk 2.

In tabel 3.1 kan de samenvatting van deze extra data worden teruggevonden.

Toestel	Clockspeed [Mhz]	Price [$]	Power [W]
PC	3250	981	79,9
Pi	1200	41,5	3,7
Nano	1479	99	10
Coral	1500	149, 99	2,65

Tabel 3.1: Gegevens voor verscheidene toestellen.

subprogramma	index	P1	P2	P3	Y1	Y2
compair	464	0	1	0.8	7	8.4
friction	14	-3			-0,29148	
narendra4	80	-0,54404			-0,45803	
pt2	208	-7,96923			-0,44761	

Tabel 3.2: Voorbeelden van de gebruikte data voor regressiemodellen.

3.3 Structuur programma

In deze sectie wordt de structuur van het hoofdprogramma besproken. De belangrijkste onderdelen van de programma's worden er toegelicht. Zo wordt er weergegeven hoe de verschillende NN-modellen opgebouwd zijn en op welke data ze worden toegepast voor zowel het trainen als het uitvoeren. De benchmark bevat in totaal 10 verschillende subprogramma's. Elk van deze is een neuraal netwerk met een zekere complexiteit bedoeld voor wijde variatie aan applicaties. Van de 10 subprogramma's kunnen er zes gecategoriseerd worden als regressie en vier als classificatie. Voor elk subprogramma wordt er uitgelegd hoe het model wordt opgesteld, hoe het getraind wordt en hoe het uiteindelijk uitgevoerd wordt. Voor de benchmark is vooral het uitvoeren van de modellen van belang. Het opstellen en trainen van een NN is een eenmalige taak en wordt bijgevolg in de praktijk niet op edge-devices gerealiseerd.

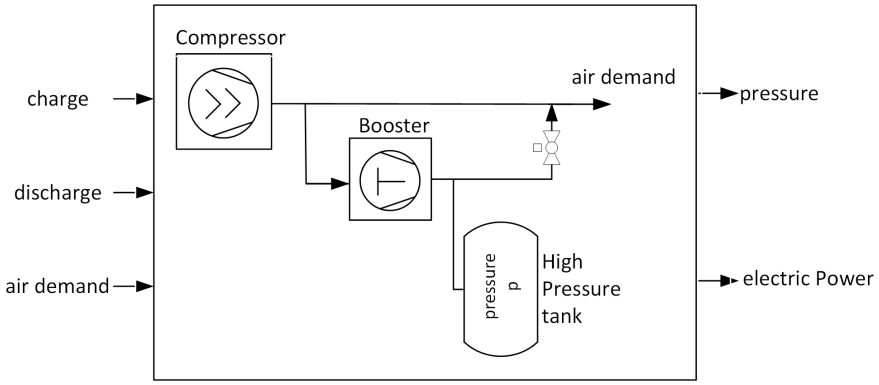
3.3.1 Regressie subprogramma's

Voor de regressie subprogramma's werd er gekozen om gebruik te maken van pyrenn. Dit is een toolbox voor zowel Python als Matlab. Deze laat op een heel eenvoudige manier NN trainen en uitvoeren. De volgende subprogramma's worden opgesteld met behulp van de pyrenn-voorbeelden.

Programma 1: compair

Het eerste subprogramma is een *compressed air storage system* of een samengedrukte lucht opslagsysteem. Het systeem heeft drie verschillende inputs en 2 gewenste outputs. De praktische werking wordt verduidelijkt in figuur 3.1 maar wordt in deze thesis niet verder op in gegaan. Hier wordt er een Recurrent Neural Network (RNN) toegepast. Dit is een regulier NN waar er een terugkoppeling bestaat tussen een node naar een vorige laag toe.

Aanmaken en trainen model Voor dit subprogramma werd er gewerkt met een dataset voorzien door Pyrenn zelf. Deze dataset levert in totaal 960 data inzendingen voor de inputs en outputs. Hiervan zijn er 480 inzendingen voorzien voor trainen en 480 voor testen van het model. In tabel 3.2 kan u een voorbeeld vinden van één datalijn. Hierbij is de inputdata P een lijst van drie features



Figuur 3.1: Praktische betekenis van het compair-subprogramma[35].

$P1, P2$ en $P3$. Analoog geldt dat de te verwachten outputdata Y een lijst voorstelt met twee features $Y1$ en $Y2$. Het NN wordt hier gedefinieerd door vier lagen. Een inputlaag, twee verborgen lagen en een outputlaag. Het aantal nodes voor de input- en outputlaag zijn gekend: drie en twee nodes respectievelijk. Voor de twee hidden layers werd er gekozen voor vijf nodes elk te implementeren. Het model kan gecreëerd worden met het commando *CreateNN()* zoals te zien is in listing 3.1. De variabele *net* bevat de vorm van het model. In het commando kunnen parameters toegevoegd worden. Zo wordt in een lijst de grootte en de lengte van de laag meegegeven. De parameters dIn, $dIntern$ en $dOut$ kunnen gebruikt worden om wederkerende verbindingen aan te maken. Zo wordt in dit subprogramma $dOut$ op waarde 1 gezet om van de outputlaag een verbinding met een vertraging van 1 tijdsperiode naar de vorige laag aan te brengen. Vervolgens wordt het model getraind met de data met het commando *train_LM()*. Hierbij worden parameters zoals k_max en E_stop toegepast om respectievelijk aan te duiden voor hoeveel iteraties er maximaal getraind mag worden en de minimale fout dat mag bereikt worden. Tot slot word het model ook opgeslagen in een Comma Seperated Value (CSV)-bestand via het commando *saveNN()*. Het uitvoeren van het model kan dan op een apart device gebeuren.

```
# Create and train NN
net = pyrenn.CreateNN([3, 5, 5, 2], dIn=[0], dIntern=[], dOut=[1])
net = pyrenn.train_LM(P, Y, net, verbose=True, k_max=500, E_stop=1e-5)
# Save outputs to certain file
prn.saveNN(net, "./models/compair.csv")
```

Listing 3.1: Creëren en trainen van pyrenn-model.

Uitvoeren model Via het commando *loadNN()* kan het model van uit een bestand terug in een variabele worden opgeslagen. Het uitvoeren van het model op testdata kan gebeuren via de instructie *NNOut()*. Het resultaat hiervan wordt in de variabele y opgeslagen zoals in listing 3.2. In vele toepassingen is het wenselijk dat variabele y zo nauw mogelijk aansluit met de echte waarden Y. In deze thesis is de accuraatheid van het model echter niet van belang. De parameters die hier onderzocht worden zijn onafhankelijk van de accuraatheid van het model. Deze worden dus ook niet berekend en verder gebruikt.

```
# Load saved NN from file
net = prn.loadNN("./models/compair.csv")
# Calculate outputs of the trained NN for train and test data
y = prn.NNOut(P, net)
```

Listing 3.2: uitvoeren van pyrenn-model.

Programma 2: friction

Het friction-subprogramma is een voorbeeld dat een fysische grootheid berekent. Het gaat hier over de wrijvingskracht F in functie van de snelheid v. Deze grootheden voldoen aan formule 3.1.

$$F = \frac{\tanh(25 \cdot v) - \tanh(v)}{2} + \frac{\tanh(v)}{5} + 0.03 \cdot v \tag{3.1}$$

Uit deze formule kan er afgeleid worden dat we met een statisch systeem met één input, v, en één output, F werken. Voor analogie met de andere pyrenn-subprogramma's worden deze respectievelijk P en Y genoemd. De pyrenn-dataset waar we hier van gebruik maken bestaat uit 41 datapunten voor het trainen en 201 datapunten voor het testen van het model. Een voorbeeld van een datapunt kan in tabel 3.2 gevonden worden. Het model dat hier gebruikt wordt is een regulier NN en bestaat uit vier lagen. De input- en outputlaag bestaan uit één node. De twee hidden layers bestaan hier elk uit drie nodes. Zowel het creëren en trainen als het uitvoeren van het model gebeuren aan analoge wijze als in listing 3.1 en 3.2.

Programma 3: narendra4

Narendra4 is een programma dat de narendra4-functie[36] beschrijft. Dit is een voorbeeld van een dynamisch systeem met slechts één output en één input met vertraging en wordt beschreven in vergelijking 3.2. Een datapunt kan gevonden worden in tabel 3.2. Het model zal ook een RNN vormen. Hier zullen er grotere terugkoppelingen aanwezig zijn. Om een output y_{k+1} te berekenen moeten de twee vorige inputs p_{k-1} en p_k ook bekend zijn naast de huidige input. Er zal dus een vertraging van twee tijdsperiodes aanwezig zijn voor de inputnode. Dit vertaalt zich in de inputvariabele dIn uit listing 3.3 die nu gelijk is aan de waarde $[1,2]$. Op analoge wijze zijn er drie tijdsperiodes vertraging aanwezig voor de outputnode: $dOut$ is nu gelijk aan de waarde $[1,2,3]$. De twee tussenliggende verborgen lagen, die elk uit drie nodes bestaan, ondervinden zelf geen vertragingen. Het uitvoeren van het RNN gebeurt weer op analoge wijze als in listing 3.2.

$$y_{k+1} = \frac{y_k \cdot y_{k-1} \cdot y_{k-2} \cdot p_{k-1} \cdot (y_{k-2} - 1) + p_k}{1 + (y_{k-1})^2 + (y_{k-2})^2} \tag{3.2}$$

```
# Create and train NN
net = pyrenn.CreateNN([1, 3, 3, 1], dIn=[1, 2], dIntern=[], dOut=[1, 2, 3])
net = pyrenn.train_LM(P, Y, net, verbose=True, k_max=200, E_stop=1e-3)
# Save outputs to certain file
prn.saveNN(net, "./models/narendra4.csv")
```

Listing 3.3: Creëren en trainen van pyrenn-model voor narendra4.

Programma 4: pt2

Het subprogramma pt2 is een programma dat een dynamisch systeem met één input en één output beschrijft. Het te gebruiken systeem hier is een tweede order transfer functie zoals in vergelijking 3.3 is opgetekend. De gebruikte pyrenn-dataset is ook hier een set met één input feature, P, en één output feature, Y. Ook van deze set is een datapunt opgenomen in tabel 3.2. In totaal zijn er 1000 datapunten beschikbaar, waarvan 500 voor het trainen en 500 voor het testen. Voor het creëren van dit model is er gekozen om naast de input- en outputlaag, twee hidden layers te implementeren met elk twee nodes. Voor deze hidden layers wordt er een vertraging van 1 tijdsperiode voorzien. Voor de uitgang wordt er een terugkoppeling van één en twee tijdsperiodes voorzien. De waardes voor $dIntern$ en $dOut$ zijn dus respectievelijk $[1]$ en $[1,2]$ bij het aanmaken van dit model. Zowel trainen en runnen gebeuren analoog aan listing 3.1 en 3.2.

$$G(s) = \frac{Y(s)}{U(s)} = \frac{10}{0.1 \cdot s^2 + s + 100} \tag{3.3}$$

Programma 5: P0Y0-narendra4

Het P0Y0-narendra4-subprogramma is een programma dat gebruik maakt van al gekende data bij het uitvoeren van een getraind netwerk. Bij een RNN is dit een interessant gegeven voor het model. Het kan meteen de vertraagde inputs and outputs een waarde geven in plaats van deze te initialiseren op nul. Dit bevordert de accuraatheid bij de start van het uitvoeren. Dit programma wordt toegepast op de narendra4-dataset. Het model wordt dus op dezelfde wijze gecreëerd en getraind. Het verschil ligt bij het uitvoeren van het model. Hierbij worden er aan het $NNOut()$ commando drie willekeurig opeenvolgende datapunten in lijstvorm gegeven voor zowel de input als output.

Programma 6: gradient

Dit subprogramma berekent de gradiënt-vector van de foutmarge van een NN. Deze berekening is mogelijk met twee verschillende algoritmen: Real Time Recurrent Learning (RTRL) en Back Propagation Trough Time (BPTT). In deze thesis wordt er gebruik gemaakt van het RTRL-algoritme. Deze werd in de documentatie beschreven als een snellere oplossing bij het uitvoeren van het model. Dit subprogramma wordt toegepast op de pt2-dataset. Het model wordt bijgevolg op dezelfde wijze gedeclareerd als het pt2-subprogramma. De train- en run-commando's zijn te vinden in listing 3.4.

```
# Create and train NN
net = prn.CreateNN([1, 2, 2, 1], dIn=[0], dIntern=[1], dOut=[1, 2])
data, net = prn.prepare_data(P, Y, net)
# Run NN
J, E, e = prn.RTRL(net, data)
```

Listing 3.4: Creëren, trainen en runnen van pyrenn-model voor gradient.

3.3.2 Classificatie subprogramma's

Programma 7: FashionMNIST

Het FashionMNIST-subprogramma is samen met NumberMNIST een van de klassiekers voor starters die kennis met ML en NN willen maken. Bovendien worden beide programma's ook regelmatig in andere benchmarks gebruikt wat vergelijkbaarheid bevordert. Voor deze redenen zullen we beiden ook in de benchmark opnemen. FashionMNIST is een NN dat foto's van kledij-stukken probeert te classificeren volgens tien mogelijke labels.

Aanmaken en trainen model Voor we het model beschrijven worden eerst de te gebruiken data verkend. De dataset[3] van foto's en labels die voor het trainen gebruikt wordt, bestaat uit 60.000 instanties. Elke instantie uit de foto-dataset omvat een foto van 28 bij 28 pixels. Elke pixel bestaat hier uit één waarde en is dus geen RGB-pixel met drie waardes. In figuur 3.2 zijn er een aantal voorbeelden van instanties terug te vinden. Voor het model opgesteld kan worden moeten de data eerst nog verwerkt worden naar een schaal die voor de compiler van het model beter te verwerken is. De waarde van één pixel varieert tussen nul en 255. Deze worden door het maximum, 255, gedeeld zodat deze tussen nul en één komen te liggen.
Vervolgens kan het model gedeclareerd worden. In listing 3.5 wordt de declaratie, compilatie en het trainen van het model getoond. Om het model op te bouwen, werd er gebruik gemaakt van Keras. Dit is een *high level interface* die meerdere deep learning libraries kan aanspreken. Het model bestaat uit drie lagen. Aan de inputlaag wordt de verwerkte data ingegeven in matrixvorm. Vervolgens worden de 28 x 28 of 784 waarden omgezet via een hidden layer met 128 nodes en een relu-activatiefunctie naar de output. In de outputlaag wordt er de $softmax$-activatiefunctie toegepast. Deze functie zorgt voor probabilistische uitkomst voor elke outputnode. Elke node zal hierdoor een waarde krijgen die overeenstemt met de kans die het model acht aan de input om overeen te komen met een bepaald label. De som van de waarden in alle outputnodes moet gelijk zijn aan één doordat enkel de 10 gebruikte labels legitieme oplossingen zijn voor het netwerk. Na het declareren, wordt het model gecompileerd. In het $compile()$-commando worden verschillende

[3] Datasets te vinden op: https://www.kaggle.com/zalando-research/fashionmnist, website geraadpleegd op 13/04/20.

Figuur 3.2: Enkele voorbeelden uit de FashionMNIST-dataset[37].

parameters zoals optimizer en loss-functie meegedeeld aan de compiler. Deze bepalen de wijze waarop het model gecompileerd wordt. Tot slot wordt met het $fit()$-commando het trainen gestart. Hier worden de verwerkte inputdata en bijhorende labels aan toegevoegd. De volgende stap is het omzetten van het getrainde model naar een TFLite-model. Deze omzetting wordt in paragraaf 3.3.3 in detail uitgelegd. Na de conversie kan het model gebruikt worden om op data toegepast te worden.

```
# Building the model
model = tf.keras.Sequential([
       keras.layers.Flatten(input_shape=(28, 28)),
       keras.layers.Dense(128, activation="relu"),
       # the probability for each given class (total =1)
       keras.layers.Dense(10, activation="softmax")])
# Compile the model
model.compile(optimizer="adam",
       loss="sparse_categorical_crossentropy",
       metrics=["accuracy"])
# training the model
model.fit(train_images, train_labels, epochs=5)
```

Listing 3.5: Creëren en trainen van sequentieel model voor FashionMNIST.

Uitvoeren model Het uitvoeren van een TFLite-model gebeurt op een andere wijze dan een standaard TF-model. Bij een gewoon model wordt de $predict()$-methode toegepast. Bij een TFLite-model wordt er eerst een *interpreter* gedeclareerd waar de verschillende tensors aan gealloceerd worden. Vervolgens kan de ingangstensor toegewezen worden met de $set\_tensor()$-methode. Daarna kan het model gerund worden op de ingangstensor, waarna de output naar de output-tensor wordt gestuurd. Met de $get\_tensor()$ kan de output opgehaald en verwerkt worden. Dit is terug te vinden in listing 3.6.

```
# Load TFLite model and allocate tensors.
interpreter = tf.lite.Interpreter(model_path=path_model)
interpreter.allocate_tensors()
# Run TFLite model
interpreter.set_tensor(input_details[0]['index'], input_data)
interpreter.invoke()
output_data = interpreter.get_tensor(output_details[0]['index'])
```

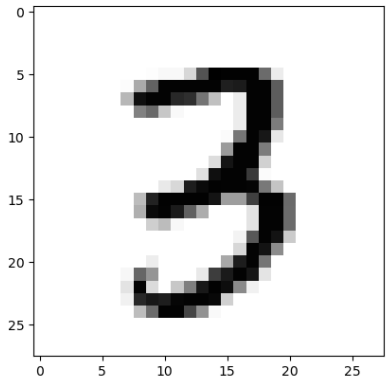
Listing 3.6: Runnen van sequentieel model voor FashionMNIST.

Programma 8: NumberMNIST

NumberMNIST is net zoals FashionMNIST een gekend voorbeeld in de ML-wereld. In dit programma wordt er een model opgesteld met als doel handgeschreven getallen van nul t.e.m. negen te herkennen. Door de eenvoud is het mogelijk om dit programma op een volledig analoge wijze te verwezenlijken zoals in FashionMNIST. Om te vermijden dat dit een exacte kopie wordt, is er voor gekozen om gebruik te maken van een ander type model dan in FashionMNIST.

Aanmaken en trainen model De te gebruiken dataset[4] voor dit model vertoont vele gelijkenissen met de dataset van FashionMNIST. Elk datapunt bestaat uit een foto van 28 bij 28 pixels. Elke pixel bestaat uit slechts één waarde i.p.v. drie zoals bij een RGB-afbeelding. De train-dataset bestaat 60.000 afbeeldingen, de test-dataset uit 10.000. Een voorbeeld van een afbeelding is te vinden in figuur 3.3. De data worden voor dit programma op dezelfde manier voor verwerkt als in FashionMNIST. De grootte van de pixels wordt door de waarde 255 gedeeld zodat de waarde van de pixels tussen nul en één liggen. Voor het opstellen van het model wordt er een andere richting uit gegaan. Voor dit model wordt er een CNN opgebouwd. De structuur ervan kan in listing 3.7 teruggevonden worden. De $Conv2D()$-methode zorgt voor de herkenning van bepaalde vormen in de afbeelding ongeacht de plaats. De opeenvolgende lagen brengen de vorm op een bepaalde plaats in verband met het juiste getal. De outputlaag is een laag met 10 nodes, één voor elk getal, waar opnieuw een $softmax$-functie op toegepast wordt. Het compilen en het trainen van het model gebeurt analoog aan FashionMNIST zoals in listing 3.5.

[4] Datasets te vinden op: https://www.kaggle.com/c/digit-recognizer/data, website geraadpleegd op 13/04/20.



Figuur 3.3: Een voorbeeld uit de NumberMNIST-dataset.

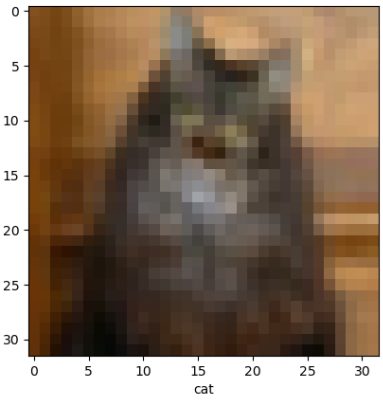
```
# Creating a Sequential Model and adding the layers
model = keras.models.Sequential()
model.add(keras.layers.Conv2D(28, kernel_size=(3, 3), input_shape=input_shape))
model.add(keras.layers.MaxPooling2D(pool_size=(2, 2)))
model.add(keras.layers.Flatten()) # Flattening the 2D arrays
model.add(keras.layers.Dense(128, activation=tf.nn.relu))
model.add(keras.layers.Dropout(0.2))
model.add(keras.layers.Dense(10, activation=tf.nn.softmax))
```

Listing 3.7: Structuur van het Convolutioneel Neuraal Netwerk NumberMNIST.

Uitvoeren model Ook het uitvoeren van het model gebeurt op gelijkaardige wijze aan de methode in FashionMNIST. Er wordt een interpreter aangemaakt vanuit een opgeslagen model. Aan deze interpreter worden tensors toegekend die vervolgens worden ingevuld met testdata. Deze worden door het $invoke()$ commando uitgevoerd en resulteren in een tensor met de outputresultaten.

Programma 9: catsVSdogs

CatsVSdogs is het derde programma dat berust op classificatie. Het is een programma dat de inhoud van een afbeelding kan herkennen en onderverdelen volgens twee klassen: *cat* of *dog*. Dit programma is een verderzetting van een algoritme dat in staat is om tien verschillende voorwerpen te herkennen. Door de toevoeging van de laatste laag $Dense(2, activation = "softmax")$, te zien in listing 3.8, en een aanpassing van de datalabels is het mogelijk het model te vereenvoudigen naar herkennen van enkel katten en honden.



Figuur 3.4: Een voorbeeld van een kat uit de catsVSdogs-dataset.

De dataset die hier werd toegepast is de cifar10-dataset[5]. Deze bevat 60.000 afbeeldingen van 32 bij 32 pixels. Hiervan worden er standaard 50.000 afbeeldingen gebruikt worden voor het trainen van het convolutioneel model en de overige voor het testen hiervan. In deze thesis wordt er echter voor gekozen om het aantal testafbeeldingen te verlagen naar 80 afbeeldingen van katten en honden. Dit wordt gedaan om sterk uiteenlopende looptijden in de resultaten te vermijden. In figuur 3.4 is een afbeelding opgenomen. Hier is te zien dat een pixel, een RGB-pixel is met een waarde voor elke kleur. De vorm van de input is hier dus een list van $[32, 32, 3]$. Ook in dit subprogramma worden de data voorverwerkt door de grootte van de pixelwaarden te reduceren tot een getal tussen nul en één. Dit wordt gerealiseerd door alle waardes te delen door 255. De volgende stappen, zoals compileren, trainen en runnen van het model, kan nu op analoge wijze gebeuren zoals aangegeven in paragraaf 3.3.2.

```
# Creating a Sequential Model and adding the layers
model = models.Sequential()
model.add(layers.Conv2D(32, (3, 3), activation='relu', input_shape=(32, 32, 3)))
model.add(layers.MaxPooling2D((2, 2)))
model.add(layers.Conv2D(64, (3, 3), activation='relu'))
model.add(layers.MaxPooling2D((2, 2)))
model.add(layers.Conv2D(64, (3, 3), activation='relu'))
model.add(layers.Flatten())
model.add(layers.Dense(64, activation='relu'))
model.add(layers.Dense(10, activation='relu'))
model.add(layers.Dense(2, activation="softmax"))
```

Listing 3.8: Structuur van het Convolutioneel Neuraal Netwerk catsVSdogs.

[5]Datasets te vinden op: https://www.cs.toronto.edu/ kriz/cifar.html, website geraadpleegd op 15/04/20.

Figuur 3.5: Een voorbeeld uit de Image Recognition-dataset.

Programma 10: Image Recognition

Het laatste subprogramma dat valt onder classificatie is het Image Recognition-programma. Dit programma is in staat om 1001 verschillende voorwerpen te herkennen op alledaagse foto's. In figuur 3.5 is een voorbeeld van zo'n afbeelding te vinden. Deze grote taak wordt verwezenlijkt door te steunen op een model dat door Google werd ontwikkeld: Mobilenet[6]. Dit is een model getraind op enkele miljoenen afbeeldingen, en bestaat uit meer dan 80 lagen en 3,2 miljoen trainbare parameters. Door gebruik te maken van dit uitgebreide model zal er geen nood zijn aan het zelf compilen of trainen van een model. De data die op dit model toegepast worden, zijn afbeeldingen van 224 bij 224 pixels. Deze data worden verwerkt maar deze keer niet tussen nul en één. De range van de data wordt herleid tot het bereik $[-1, 1]$. Voor het uitvoeren van het model worden er 124 afbeeldingen voorzien.

3.3.3 Conversie naar TFLite

Doordat er in deze thesis gebruik gemaakt wordt van devices die enkel met TFLite werken worden de modellen omgezet naar een TFLite-model. Dit gebeurt door een convertor-object te creëren van het bestaande keras-model. Op deze convertor worden optimalisatietechnieken uitgevoerd waarna het model wordt omgezet naar een TFLite-equivalent model. Dit model wordt vervolgens met het commando $write()$ uitgeschreven naar het gewenste bestand. Dit nieuwe model kan vervolgens opgeroepen worden voor uitvoering op gewenste tijdstippen. In listing 3.9 kan de gebruikte code terug gevonden worden.

```
# Convert the model
converter = tf.lite.TFLiteConverter.from_keras_model(model)
converter.optimizations = [tf.lite.Optimize.DEFAULT]
tflite_quant_model = converter.convert()
# Saving tflite model
open(path_model + "fashionMNISTmodel.tflite", "wb").write(tflite_quant_model)
```

Listing 3.9: Converteren naar een TFLite-model.

[6] Datasets te vinden op: https://ai.googleblog.com/2017/06/mobilenets-open-source-models-for.html, website geraadpleegd op 15/04/20.

Als het model gebruiksklaar is voor TFLite, wordt van het programma ook een kopie gemaakt die op de TFLite-devices zoals de Coral Dev Board uitgevoerd kan worden. Er wordt dus gewerkt met twee programma's die heel weinig van elkaar verschillen. Ze voeren dezelfde programma's uit op dezelfde wijze. De originele versie werkt voor de classificatieprogramma's met de standaard tensorflow-libraries waar TFLite een onderdeel van is. De aangepaste versie werkt op een stand-alone versie van de TFLite-module. Deze laatste bibliotheek werkt enkel op speciaal ontwikkelde hardware zoals de Coral Dev Board. De aan te brengen veranderingen zijn terug te vinden in listing 3.10.

Het betreft twee belangrijke aanpassingen. De verandering van module en toevoeging van het $libedgetpu.so.1$-bestand. Dit bestand is de Edge TPU runtime library. Deze bibliotheek helpt bij het verdelen van instructies over de CPU en TPU.

```
# Lines in original File:
import tensorflow as tf
interpreter = tf.lite.Interpreter(model_path=path_model)
# Lines in TFLite-compatible File:
import tflite_runtime.interpreter as tflite
interpreter = tflite.Interpreter(model_path=path_model,
                                 experimental_delegates=[tflite.load_delegate('libedgetpu.so.1')])
```

Listing 3.10: Converteren naar een TFLite-programma.

3.4 Uitvoeren metingen

In de benchmark worden twee belangrijke parameters gemeten. De tijdsduur dat het programma over het uitvoeren doet en het percentage van de CPU dat gebruikt wordt tijdens het uitvoeren. Voor het meten van de tijdsduur wordt er gebruik gemaakt van de $time$-module. Met behulp van deze module kan de tijd via het commando $time.time()$ uitgedrukt worden in een floating point getal uitgedrukt in seconden. Als de tijd bij aanvang van het uitvoeren van het model opgeslagen wordt alsook bij het stoppen van het model, dan kan men de tijdsduur bekomen door het verschil te nemen tussen de stop-waarde en de start-waarde.
Om het gebruik van de CPU te meten wordt er gebruik gemaakt van de psutil-module. Deze module bevat het commando $psutil.cpu\_percent()$ waarmee het verbruik in percent kan opgevraagd worden. Het opvragen van dit commando geeft het verbruik weer, sinds de laatste keer dat het commando werd opgevraagd. Dit betekent dat de functie twee keer opgeroepen moet worden. De eerste keer vlak voor het starten van het uitvoeren van het programma, de tweede keer vlak erna. Als de functie voor het eerst opgeroepen wordt, zal de waarde niet opgeslagen worden. De waarde die teruggegeven wordt bij de tweede keer opvragen wordt wel opgeslagen. Deze bevat het juiste percentage van het verbruik van de CPU sinds de start van het uitvoeren van het programma. Met de parameter $percpu$ kan er per core in de CPU gemeten worden. Aangezien alle te gebruiken devices over meerdere kernen beschikken zetten wij deze op $True$ om een beter beeld te krijgen van de werking van het programma. In listing 3.11 bevindt zich een voorbeeld van de meetmethode in pseudocode.
Om statistisch representatieve resultaten te bekomen zullen de metingen meerdere keren gebeuren. In deze thesis wordt er voor gekozen om elk programma 20 iteraties te laten voltooien. Naast de metingen van elk programma wordt er ook nog een meting gedaan over de totale duur van het

uitvoeren van de verschillende programma's en het verbruik van de CPU in idle-toestand.

```
# Eerste keer opvragen van tijd en CPU-verbruik
psutil.cpu_percent(interval=None, percpu=True)
time_start = time.time()
# Uitvoeren van het model op testdata
model.run(y_test)
# Tweede keer opvragen van tijd en CPU-verbruik en opslag in gewenste variabelen
time_stop = time.time()
cpu_data = psutil.cpu_percent(interval=None, percpu=True)
time_run = time_stop - time_start
time_total += time_run
```

Listing 3.11: Meten van gewenste data.

3.5 Opslaan van data

Een belangrijk onderdeel van deze benchmark is het opslaan van de gegenereerde data. Om te voorkomen dat de data verwarrend en ongestructureerd is, wordt er getracht om data op gestructureerde wijze op te slaan in een bestand. Deze kan dan later op eenvoudige wijze verwerkt en gevisualiseerd worden.
Voor de data opgeslagen worden, gebeurt er een controle naar de waarde van de data. Indien het zou blijken dat tijdens het meten onrealistische waarden gegenereerd worden voor een bepaalde iteratie (bijvoorbeeld 0% CPU-verbruik) dan wordt de iteratie opnieuw uitgevoerd. Indien dit niet het geval is, worden de data gelogd naar een CSV-bestand. Het bestand krijgt een unieke bestandsnaam gelinkt aan het moment van uitvoeren en aan het toestel waar de code op uitgevoerd wordt. In listing 3.12 kan de log-functie teruggevonden worden.

```
def logging_data(program_index, stop, start, cpu):
# Logging data
cores_avg = mean(cpu)
time_diff = stop-start
with open("unique_file_name.csv", mode='a+') as data_file:
       data_writer = csv.DictWriter(data_file, fieldnames=fieldnames)
       data_writer.writerow(
              {'Naam': labels[program_index],
              'CPU Percentage': str(cores_avg),
              'timediff': str(time_diff)})
```

Listing 3.12: Opslaan van de gewenste data.

Hoofdstuk 4

Data verwerking

In dit hoofdstuk wordt de data verkregen uit vorig hoofdstuk verwerkt in visuele resultaten. De verschillende ondernomen stappen worden besproken. De eerste stap behandelt het verwerpen van bepaalde datapunten. Vervolgens wordt de data met bepaalde parameters in verband gebracht. En tot slot wordt de data op vlak van spreiding geanalyseerd en uiteindelijk in grafiekvorm weergegeven. De resultaten zelf worden in hoofdstuk 5 besproken.

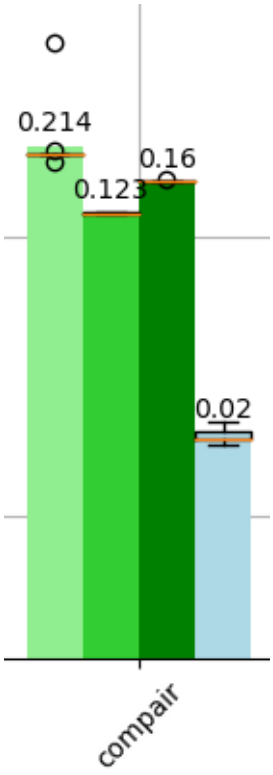
4.1 Data cleaning

Een belangrijk gebeuren voor de data verwerkt kan worden, is het schoonmaken van de data. Tijdens het uitvoeren van metingen is het mogelijk dat er meetfouten plaatsvinden. Deze kunnen door allerlei oorzaken plaats vinden en moeten zo veel mogelijk vermeden worden. In deze sectie bespreken we een aantal van deze meetfouten en hoe deze aangepakt en verwerkt worden.

4.1.1 Verwerpen data

Een eerste wijze van schoonmaken is het verwerpen van datapunten waarvan er geweten is dat deze onmogelijk correct kunnen zijn. Dit betreft meetfouten waar bijvoorbeeld de duur van het uitvoeren van een programma gelijk is aan nul seconden. Dit is uiteraard niet mogelijk, elke meting heeft namelijk een duur van bepaalde grootte. Een andere meetfout die geregeld op trad kwam bij het meten van het CPU-verbruik voor. Indien na het uitvoeren van een programma bleek dat er geen activiteit in de CPU werd gemeten is dit als gevolg van een meetfout.

Indien deze meetfouten worden gedetecteerd zal de volledige meting worden verworpen en wordt de meting herhaald. De controle op deze meetfouten vindt dus plaats na het uitvoeren op de meting en voor het opslaan van de data. In listing 4.1 kan deze controle teruggevonden worden. Alleen nadat er geen meetfout gedetecteerd wordt de data uitgeschreven en wordt de iteratie erkend als een geldige iteratie. Door een imperfecte meting te hernemen blijft het totaal aantal bruikbare datapunten bijgevolg altijd 20.



Figuur 4.1: Een voorbeeld van uitschieters in data van de duur van het compairprogramma.

```
# Meting wordt gecontroleerd voor
if (mean(cores) != 0.0) and (time_stop-time_start != 0):
       logging_data(10, time_stop, time_start, cores)
       iteration += 1
print("iteration: ", iteration, " mean cores: ", mean(cores), " duration: ",
    time_stop-time_start)
```

Listing 4.1: Controleren op meetfouten.

4.1.2 Behandelen uitschieters

De tweede wijze waarop de data wordt schoon gemaakt is door het behandelen van uitschieters. Tijdens het meten is het mogelijk dat een datapunt verder of dichter van het gemiddelde licht door een externe factor. Zo kan bijvoorbeeld een subroutine van het Operating System de meting vertragen en hierdoor de meting beïnvloeden. Om deze invloeden te vermijden is het nodig om de uitschieters of outliers te herkennen en te elimineren. Outliers worden hier beschouwd als datapunten die een grotere afwijking van het gemiddelde hebben dan drie standaardafwijkingen. Om deze outliers te vinden wordt er een boxplot opgesteld met behulp van de module matplotlib. In figuur 4.1 kan een voorbeeld van data met een outlier gevonden worden. Zodra de voornaamste outliers geïdentificeerd zijn, zullen deze aangepast worden. De data wordt veranderd in het gemiddelde van de overige niet-uitschieters om het gemiddelde van de ware data niet te wijzigen.

4.2 Vormgeving resultaten

In deze sectie wordt er besproken uit welke grootheden de resultaten zullen bestaan, hoe deze gevormd worden en op welke manier de resultaten gevisualiseerd worden.

De belangrijkste te onderzoeken parameter is de tijd. De duur van uitvoeren dat een programma in beslag neemt kan een grote factor spelen bij het maken van een kosten-baten analyse. Echter als er enkel de latency met elkaar vergeleken wordt, kan dit leiden tot een vertekend beeld. Daarom worden ook factoren zoals CPU-gebruik en kloksnelheid in rekening gebracht. De latency per percentage CPU-gebruik en per MHz kloksnelheid geeft een beter beeld van de performantie van elk toestel. Een andere variabele die in rekening gebracht kan worden is het vermogen dat elk toestel verbruikt. Door de latency te vermenigvuldigen met het vermogen bekomt men de energie die verbruikt wordt tijdens de executie. Voor de grootheden geldt:

$$time \cdot power = s \cdot \frac{J}{s} = J = Energy \tag{4.1}$$

Het is interessant om de verbruikte energie van de verschillende toestellen voor hetzelfde programma met elkaar te vergelijken. Toepassingen die energie gebonden zijn of waar een batterij de voeding voorziet kunnen een afweging maken tussen het energieverbruik en de duur van executie.

4.3 Overzicht code

Hoofdstuk 5

Resultaten

Hoofdstuk 6

Conclusie

Bibliografie

[1] M. R. Minar and J. Naher, “Recent advances in deep learning: An overview,” 02 2018.

[2] J. Bloem, M. Van Doorn, S. Duivestein, D. Excoffier, R. Maas, and E. Van Ommeren, “The fourth industrial revolution,” *Things Tighten*, vol. 8, 2014.

[3] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[4] L. Barreto, A. Amaral, and T. Pereira, “Industry 4.0 implications in logistics: an overview,” *Procedia Manufacturing*, vol. 13, pp. 1245–1252, 2017.

[5] H. Li, K. Ota, and M. Dong, “Learning iot in edge: Deep learning for the internet of things with edge computing,” *IEEE Network*, vol. 32, no. 1, pp. 96–101, Jan 2018.

[6] D. I. Poole, R. G. Goebel, and A. K. Mackworth, *Computational intelligence*. Oxford University Press New York, 1998.

[7] https://www.frankwatching.com/archive/2017/02/06/artificial-intelligence-6-redenen-\
waarom-je-juist-nu-moet-starten/, accessed on 27.1.2020.

[8] https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html, accessed on 27.1.2020.

[9] J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, “An overview of machine learning,” in *Machine learning*. Springer, 1983, pp. 3–23.

[10] http://webindream.com/reinforcement-learning/, accessed on 23.12.2019.

[11] https://github.com/drewnoff/spark-notebook-ml-labs/tree/master/labs/DLFramework, accessed on 24.12.2019.

[12] https://pathmind.com/wiki/neural-network, accessed on 23.12.2019.

[13] J. Han and C. Moraga, “The influence of the sigmoid function parameters on the speed of backpropagation learning,” in *International Workshop on Artificial Neural Networks*. Springer, 1995, pp. 195–201.

[14] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[15] https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6, accessed on 23.12.2019.

[16] https://medium.com/machine-learning-bites/machine-learning-decision-tree-classifier-9eb67cad263e, accessed on 24.12.2019.

[17] F. Sun, “Kernel coherence encoders,” 2018.

[18] https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781788994170/5/ch05lvl1sec36/support-vector-machines, accessed on 20.12.2019.

[19] I. Sieben and L. Linssen, “Logistische regressie analyse: een handleiding,” *Geraadpleegd via www. ru. nl/publish/pages/525898/logistischeregressie. pdf*, 2009.

[20] https://en.wikipedia.org/wiki/Regression_analysis, accessed on 19.12.2019.

[21] https://elitedatascience.com/machine-learning-algorithms, accessed on 21.12.2019.

[22] https://nl.wikipedia.org/wiki/Singleboardcomputer, accessed on 23.12.2019.

[23] https://beagleboard.org/ai, accessed on 23.12.2019.

[24] https://venturebeat.com/2019/10/22/googles-coral-ai-edge-hardware-launches-out-of-beta/, accessed on 23.12.2019.

[25] https://developer.nvidia.com/embedded/jetson-nano, accessed on 23.12.2019.

[26] https://developer.nvidia.com/embedded/jetson-tx2, accessed on 23.12.2019.

[27] https://www.raspberrypi.org/products/raspberry-pi-4-model-b/, accessed on 23.12.2019.

[28] https://www.notebookcheck.net/Lenovo-Legion-Y520-15IKBN-7300HQ-GTX-1050-Ti-FHD-Laptop-Review.256682.0.html, accessed on 20.04.2020.

[29] L. A. Libutti, F. D. Igual, L. Pinuel, L. De Giusti, and M. Naiouf, “Benchmarking performance and power of usb accelerators for inference with mlperf.”

[30] https://devblogs.nvidia.com/jetson-nano-ai-computing/, accessed on 17.02.2020.

[31] L. P. Bordignon and A. von Wangenheim, “Benchmarking deep learning models on jetson tx2.”

[32] https://en.wikipedia.org/wiki/LINPACK_benchmarks, accessed on 27.02.2020.

[33] https://www.top500.org/project/linpack/, accessed on 27.02.2020.

[34] https://keras.io/backend/, accessed on 16.04.2020.

[35] https://pyrenn.readthedocs.io/en/latest/examples.html, accessed on 10/04/2020.

[36] K. S. Narendra and K. Parthasarathy, “Identification and control of dynamical systems using neural networks,” *IEEE Transactions on Neural Networks*, vol. 1, no. 1, pp. 4–27, 1990.

[37] https://becominghuman.ai/how-to-create-a-clothing-classifier-fashion-mnist-program-on-google-colab-99f620c24fc accessed on 13/04/2020.

Bijlage A

Een aanhangsel

sdfsffqsfsf

Bijlage B

Beschrijving van deze masterproef in de vorm van een wetenschappelijk artikel

Bijlage C

Poster

KU LEUVEN

Ghent Technology Campus
2019-2020

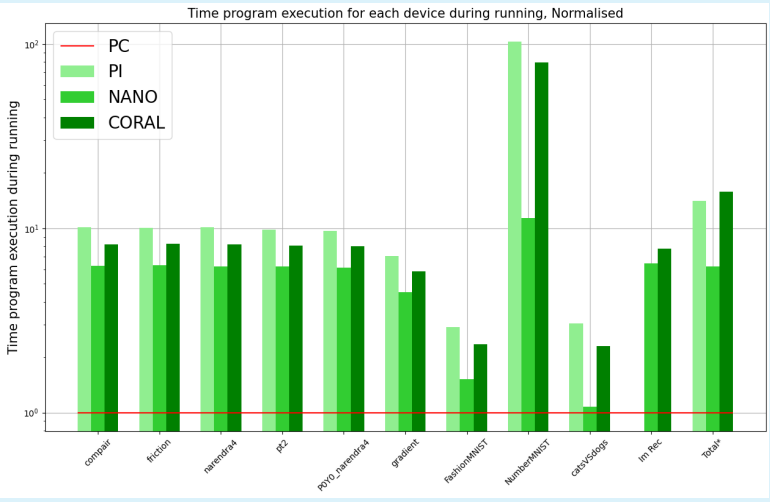
Comparative Study of Single Board Computers for AI Edge Computing Purposes

Introduction

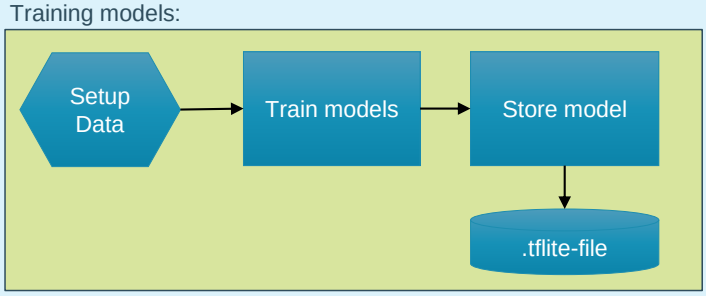
The deployment of self-driving vehicles is one of many new and interesting low-latency applications. To aid to lower latency during localization in position fixing, a benchmark is proposed to examine the potential of executing a trained Neural Networks (NN) or Machine Learning (ML) algorithm on edge devices like the Nvidia Jetson Nano, Google Coral Dev and Raspberry Pi 3 compared to regular computer.

Data acquisition

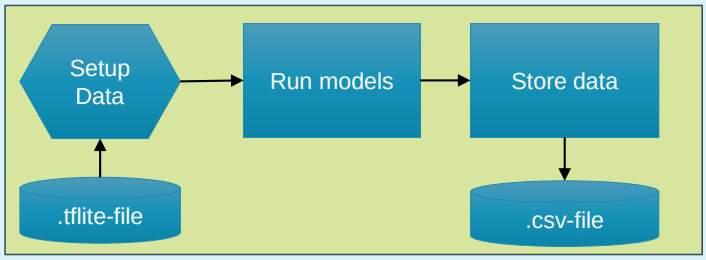
In order to achieve a representative benchmark, it is necessary to investigate the performance of different programs on each edge device in an identical manner. In this benchmark several different programs are tested spread across categories of NN as regression and classification. For each program the duration of execution and utilization of the processor unit is measured and stored. To achieve statistically results, every program will be run for 20 different iterations.



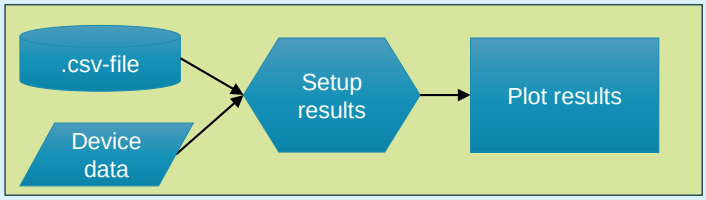
Training models:



Running models:



Plotting results:



Data analysis

To compare the performance of different edge devices, latency data will be adjusted for variables like processor usage, clock speed, price and energy usage. The results are normalised to ease a comparison.

Conclusion

The results show that among edge-devices, the Jetson Nano is the Single Board Computer with the lowest latency. When comparing power consumption for the given latency, the Coral Dev is the most power efficient.

Student: Arno Plaetinck
Promotor: Prof. Dr. Ir. Nobby Stevens
Co-promotors: Ing. Willem Raes
Ing. Jorik De Bruycker

KU Leuven
Ghent Technology Campus
Gebroeders de Smetstraat 1
9000 Ghent

Master in Electronics and ICT Engineering
Technology option Embedded Systems
Academic year: 2019-2020
Date: 6/04/2020

FACULTEIT INDUSTRIELE INGENIEURSWETENSCHAPPEN
TECHNOLOGIECAMPUS GENT
Gebroeders De Smetstraat 1
9000 GENT, België
tel. + 32 92 65 86 10
fax + 32 92 25 62 69
iiw.gent@kuleuven.be
www.iiw.kuleuven.be