Val av hårdvara

## Mål

* Vi vill undersöka olika typer av hårdvara som är lämpliga att implementera neurala nätverk på.
* Vi vill hitta den lösning som ger bäst balans mellan prestanda, enkelhet att implementera, och flexibilitet i implementationen.
* Vi är speciellt intresserade av hårdvara som passar för inference på lågenergi-hårdvara, som kan användas ”on the edge”.

## Förslag

Förslag på olika grupper av plattformar:

|  |  |  |
| --- | --- | --- |
| FPGA | ASIC | CPU |
| * Kan programmeras till det man önskar * Arbetar både parallellt och sekventiellt * Flexibel * Lågnivåspråk krävs (HDL) | * Speciellt framtagen för att vara snabb på att lösa sin uppgift * Oflexibel * Bör alltid vara snabbast och mest energieffektiv | * Generell typ av hårdvara, jobbar framför allt sekventiellt * Kan använda högnivåspråk och frameworks |

Förslag på specifik hårdvara i varje grupp:

|  |  |  |
| --- | --- | --- |
| Arty Z7-20 | Google Coral dev board | Raspberry Pi 4 |
| * CPU: Cortex-A9 dual-core 650MHz * RAM: 512 MB * FPGA: XC7Z020 (jämförbar med Artix-7)   -Slices: 13,300  -LUTs: 53,200  -Flop-Flops: 106,400  -Block RAM: 630 kB | * CPU: Cortex-A53 quad-core 1,5 GHz * RAM: 1 GB * TPU: Google Edge TPU, en ASIC för att köra NN ”on the edge” | * CPU: Broadcom quad-core A72 1,5GHz * RAM: 1/2/4 GB * GPU: Broadcom VideoCore VI |
| * Har ett SoC med CPU och FPGA * Kan både använda frameworks som översätter till HDL och implementera HDL direkt * Kan bestämma vilka delar som ska exekveras i mjukvara (CPU) och i hårdvara (FPGA) | * SoM med CPU och TPU (designad för inference). * Endast kompatibel med Tensorflow Lite-modeller | * Väl använd i utvecklingsprojekt * Billigt inköpspris * Har både CPU och GPU (dock osäker om det finns frameworks som använder RP-GPU) |

## Alternativ

Alternativ till Arty Z7 skulle vara att köra på en ren FPGA, istället för att som med Arty ha FPGA kopplad till en CPU. Den FPGA som sitter i Arty är nästintill identisk med Artix-7, en FPGA också från Xilinx. Skillnaden är att vissa pinnar på Artix som används för konfigurering/bootning nu är kopplade till CPUn. Det gör att på Arty kan inte FPGAn jobba helt fristående, utan måste bootas med hjälp av CPUn. När bootning är klart fungerar dock FPGAn på Arty som en fristående FPGA. Skillnaden i bootningsprocess kommer inte att påverka ML-prestanda. Fördelen med CPU känns då större, eftersom att man kan använda färdiga frameworks och andra språk (t.ex. python/C) vilket gör implementeringen lättare.

Alternativ till Coral dev board är antingen att ha ett en egenutvecklad ASIC, eller att försöka hitta hitta en annan färdig ASIC lösning. Eftersom att Googles är så lätt att få tag på känns det enklast och mest intressant att testa på den.

Alternativ till Raspberry Pi skulle kunna vara NVIDIA Jetson Nano. Den brukar ofta jämföras med Coral TPU eftersom de ska lösa samma problem och är i ungefär samma formfaktor, men den har ingen ASIC utan är mer som en RP med högre prestanda. Den använder GPU för accelerering och kan köra alla typer av modeller och frameworks. Den stora fördelen med RP är att den är så välkänd och kan fungera som en referens mot de andra plattformarna.

## Mjukvara

Arty Z7 kan köpas med en voucher för SDSoC, vilket möjliggör att skriva i C/C++ och att använda färdiga libraries för både mjuk- och hårdvarukomponenter. För att skriva i HDL kan Vivado WebPack användas, vilket är gratis att ladda ner.

Google Coral behöver ingen extra mjukvara. Många använder Jupyter Notebook (gratis) för att arbeta med TensorFlow.

Raspberry Pi använder en egen Linux-distro, Raspbian, som är gratis att ladda ner. Ingen extra mjukvara behövs.

## Inköp

* Arty Z7-20 inkl voucher för SDSoC
  + <https://www.digikey.se/product-detail/sv/digilent-inc/471-006/1286-1221-ND/9826690>
  + Kostnad 2148,55 kr ex. moms.
* Google Coral dev board
  + <https://www.elfa.se/sv/coral-dev-kort-seeed-studio-102110260/p/30143588?queryFromSuggest=true>
  + Kostnad 1526 kr ex. moms.
* Raspberry Pi 4/4 GB
  + <https://www.dustin.se/product/5011141621/4-model-b-4gb>
  + Kostnad 559 kr ex. moms.
* Tillbehör
  + <https://www.elfa.se/sv/usb-kabel-usb-svart-valueline-vlcp60500b10/p/30046369?track=true>
    - Micro USB-kabel till Arty och Coral. Kostnad 25,10 kr ex. moms.
  + <https://www.dustin.se/product/5011141922/okdo-raspberry-pi-multihead-power-adapter-51v-3a>
    - Nätadapter USB C, för ström till RP. Kostnad 95 kr ex. moms.
  + <https://www.dustin.se/product/5011039977/ultra-c10-uhs-u1-a1>
    - SD-kort till RP. Kostnad 55 kr ex. moms.
  + <https://www.dustin.se/product/5010588121/adapter-ac-5060hz---5v-dc-3a>
    - Ström till Arty. Kostnad 135 kr ex. moms.
  + <https://www.dustin.se/product/5011097671/natadapter-1xusb-24amp>
    - Nätadapter till Coral. Kostnad 129 kr ex. moms.
  + <https://www.dustin.se/product/5011091183/synkladdkabel-usb-c>
    - USB C-kablar. **2 st**. Kostnad 2 \* 89 kr ex. Moms (178 kr tot).

## För prisuppgift

* NVIDIA Jetson Nano
  + <https://www.elfa.se/sv/nvidia-jetson-nanotm-utvecklingssats-seeed-studio-102110268/p/30143589?queryFromSuggest=true>
  + Kostnad 1016 kr ex. moms.