Energy-efficient in-situ analytics

Cheng-Hsiang Chiu

Institutt for Informatikk
Universitetet i Tromsø – Norges arktiske universitet

Workshop on Efficient Frameworks for Compute- and Data-intensive Computing
April 25, 2019

Table of Contents

1 COAT and DAO
2 In-situ Analytics at the Edge
3 Intel Movidius Neural Compute Stick
4 Convolutional Neural Networks
5 Framework
6 Future Works

Table of Contents

1 COAT and DAO
2 In-situ Analytics at the Edge
3 Intel Movidius Neural Compute Stick
4 Convolutional Neural Networks
5 Framework
6 Future Works

COAT - 1/3

- COAT: Climate ecological Observatory for Arctic Tundra.
- A long-term, ecosystem-based and adaptive observation system.
- Aims to unravel how climate change impacts arctic tundra, and to enable prudent science-based management.

COAT - 2/3

Figure: Arctic Tundra.

(Photo credit: COAT)

- Cold and desert-like conditions.
- Average winter temperature is $-34°C$ ($-30°F$).
- Yearly precipitation, including melting snow, is 15 to 25 cm.
- Growing season ranges from 50 to 60 days.

COAT - 3/3

- Researchers carry wild life sensors, cameras and other observation devices into the field.
- Devices are manually configured in the field.
- Collected data are fetched several months later by hand.

DAO

- DAO: Distributed Arctic Observatory.
- A hardware and software solution, observation unit (OU), is proposed to settle problems of data acquirement, limited energy, and deficient network.
- OU is a configurable computer node along with a set of sensors.
- State-of-the-art: these units will autonomously monitor themselves and the environment, run software layers able to configure and run new applications, synchronize data and finally gather data to centralized servers to execute analysis processes.

Data Analytics at Centralized Servers?

- Gather data to centralized servers to execute analysis processes.
  - Pros: More computing powers, more insights, and higher accuracy.

Data Analytics at Centralized Servers?

- Gather data to centralized servers to execute analysis processes.
  - Pros: More computing powers, more insights, and higher accuracy.
  - Cons: Waste of bandwidth and energy.

Table of Contents

1 COAT and DAO
2 In-situ Analytics at the Edge
3 Intel Movidius Neural Compute Stick
4 Convolutional Neural Networks
5 Framework
6 Future Works

In-situ Analytics

- How about bringing data analytics ability to OUs?
  - Gathered data are in-situ processed.
  - Pre-processed data are transmitted back to servers for storage and/or further processing.
  - Bandwidth is preserved and energy is less consumed.

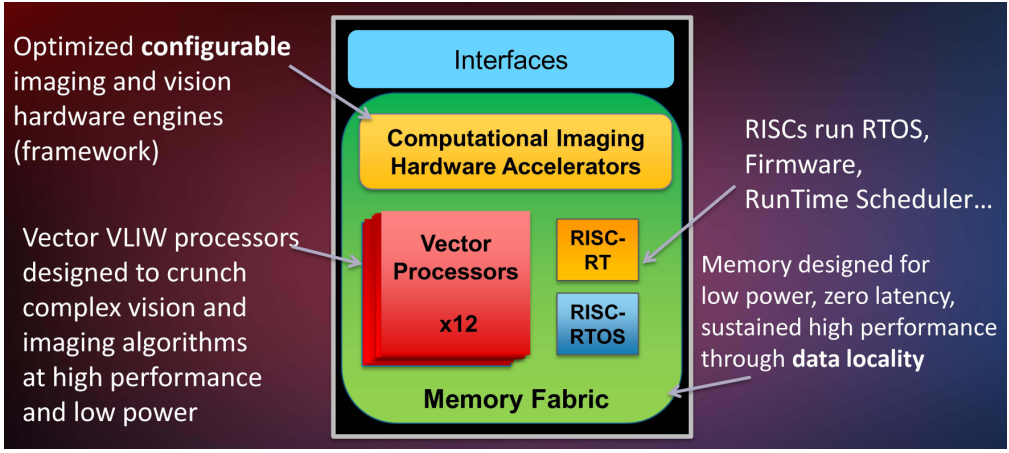
Table of Contents

Movidius Neural Compute Stick

(Photo credit: Intel)

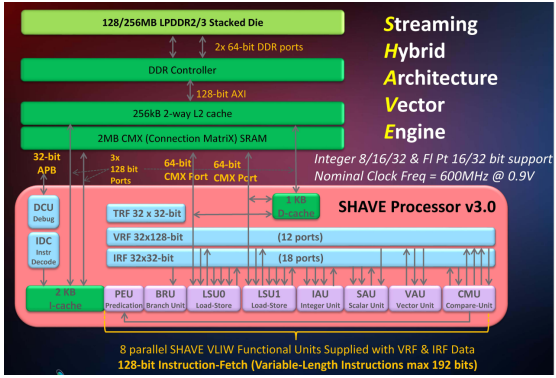
- Founded in 2005, raised $90 million and acquired by Intel in 2016.
- Designed to facilitate development, tuning and deployment of deep neural networks at the "edge" of our modern technology networks.
- Ultra low-power device.

Myriad Architecture - 1/3



(Photo credit: https://ieeexplore.ieee.org/document/7478823)

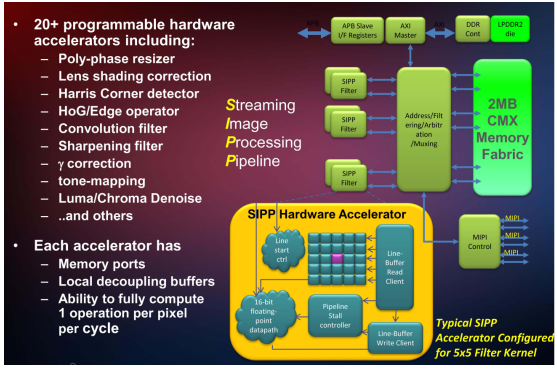
Myriad Architecture - 2/3



(Photo credit: https://ieeexplore.ieee.org/document/7478823)

- Contains wide and deep register files coupled with a Very Long Instruction Word (VLIW) for code-size efficiency.
- VLIW packets control multiple functional units which have SIMD capability for high parallelism and throughput.

Myriad Architecture - 3/3



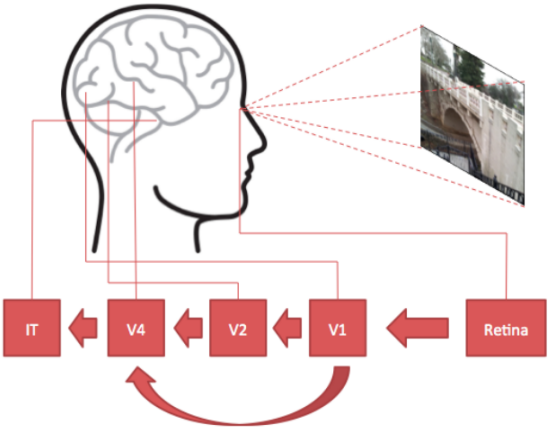
(Photo credit: https://ieeexplore.ieee.org/document/7478823)

- Provides over 20 programmable accelerators.
- Each accelerator is able to compute 1 operation per pixel per cycle.

Table of Contents

1 COAT and DAO
2 In-situ Analytics at the Edge
3 Intel Movidius Neural Compute Stick
4 Convolutional Neural Networks
5 Framework
6 Future Works

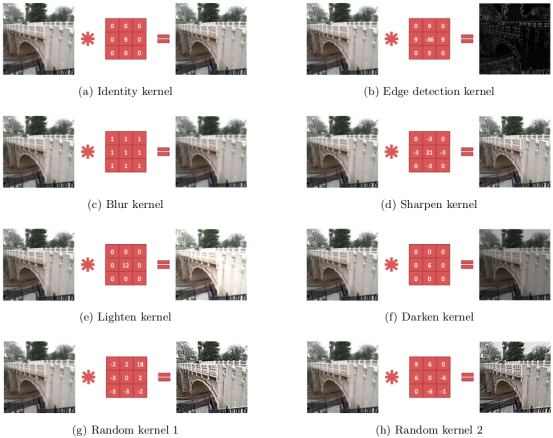
Visual Cortex



(Photo credit: Haohan Wang & Bhiksha Raj)

- Retina: Convert light energy of an object into chemical energy.
- V1: Mainly fulfills the task of edge detection.
- V2: Mainly extracts visual signals, like orientation.
- V4: Detects object features of intermediate complexity, like geometric shapes.
- Inferior temporal gyrus (IT): Performs the semantic level tasks, like face recognition.

Convoluional Operations

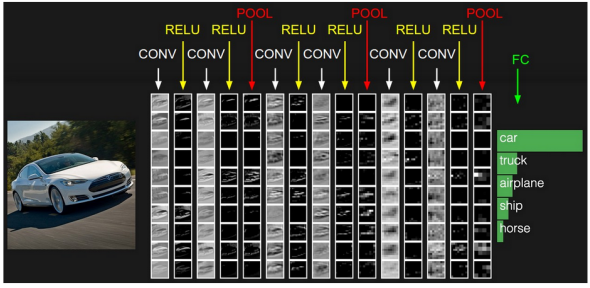


(a) Identity kernel (b) Edge detection kernel (c) Blur kernel (d) Sharpen kernel (e) Lighten kernel (f) Darken kernel (g) Random kernel 1 (h) Random kernel 2

(Photo credit: Haohan Wang & Bhiksha Raj)

- Use convolutional operations to extract features.

Convolutional Neural Networks (CNN)



(Photo credit: http://cs231n.github.io/convolutional-networks/)

- Usually consists of convolutional layers, relu layers, and pooling layers.
- LeNet, AlexNet, GoogLeNet, VGGNet, ResNet.

Table of Contents

Framework

1. Train an Inception-v3 model with COAT dataset at centralized servers.
2. Download the trained model to Movidius Neural Compute Stick.
3. Embed the stick on OUs.
4. Re-train the model if necessary.

Table of Contents

Future Works

- Parallelize CNN models.

Future Works

- Parallelize CNN models.
- Introduce temporal stream to the models for higher accuracy.
  - One stream for spatial stream, and the other stream for temporal stream.

Future Works

- Parallelize CNN models.
- Introduce temporal stream to the models for higher accuracy.
  - One stream for spatial stream, and the other stream for temporal stream.
- Apply different CNN models for different energy budgets.
  - Sacrifice accuracy for less energy consumption.

Thank you