Embedded Deep Learning

Fabio Frazao

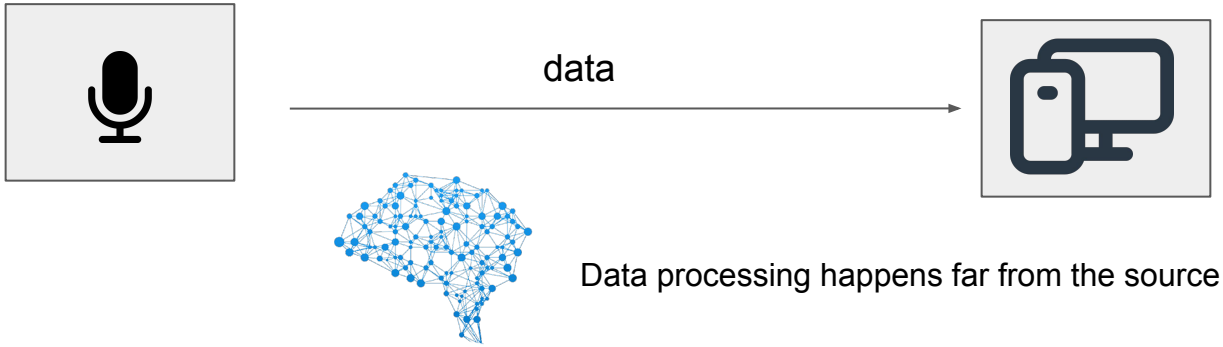
MERIDIAN, Institute for Big Data Analytics,
Dalhousie University, Halifax, Canada

Introduction

- Deep learning on the edge
- Platforms
- Constraints

Deep learning on the edge

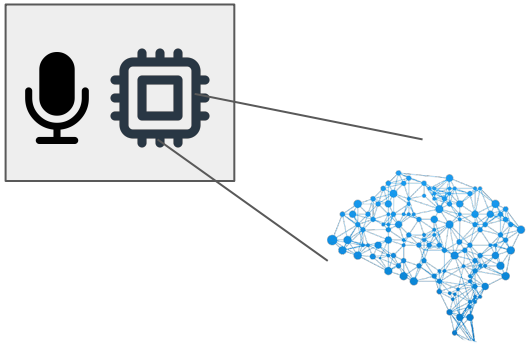
- Conventional pipeline



Data processing happens far from the source

Deep learning on the edge

- Edge computing

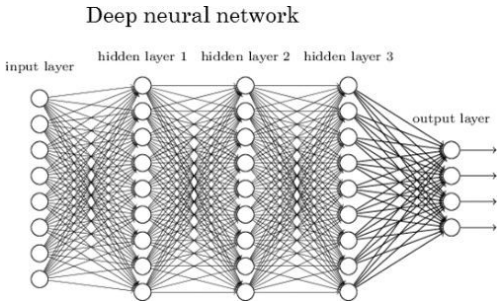


Data processing happens near the source, with the processing algorithms embedded into the hardware

Deep learning on the edge

- What is deep learning?

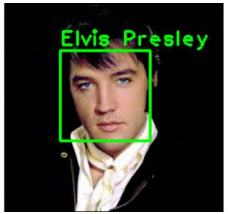
Deep learning is an approach to **machine learning** that uses **deep neural networks**



Everyday applications:

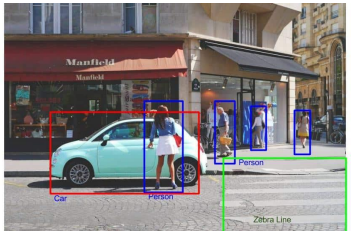
Speech recognition & synthesis

Face Recognition

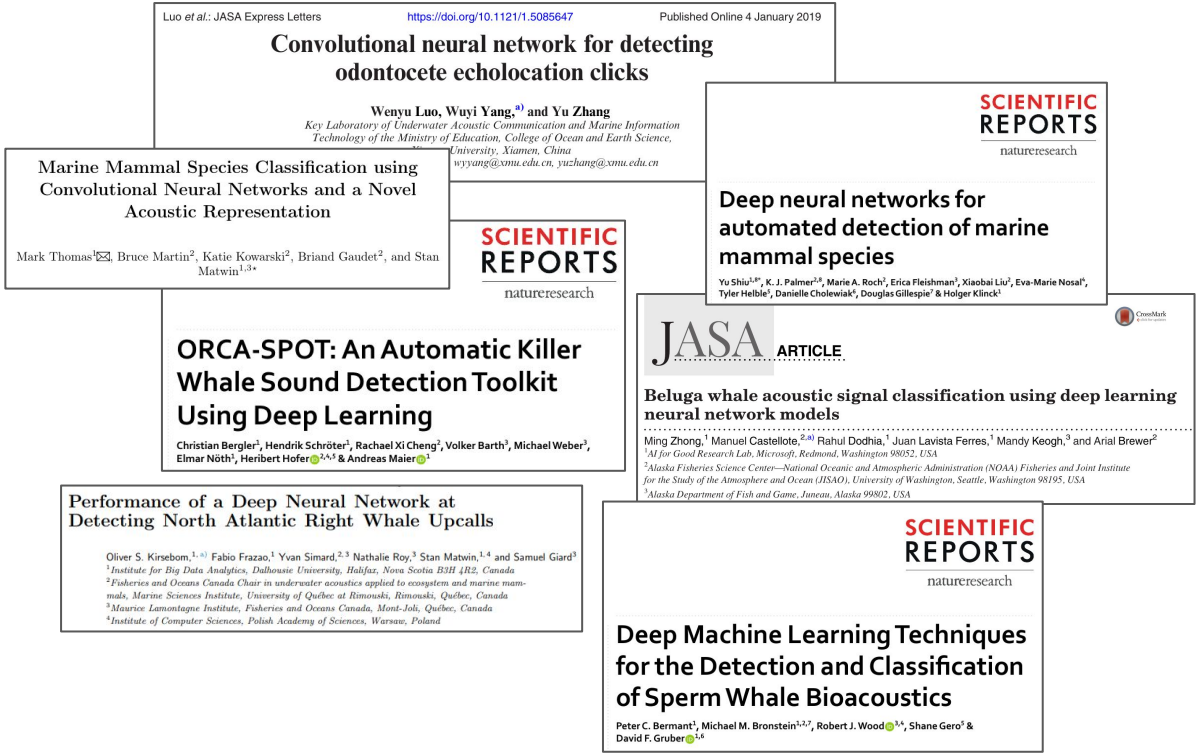


Translation

Object detection



Deep learning on the edge



Staple deep neural networks are useful for underwater acoustics.

+ Better performance in many tasks
+ High adaptability and reusability
- Requires more resources (data, computing power, etc)
- Often hard to interpret

Can they be embedded into PAM systems?

Platforms for edge computing

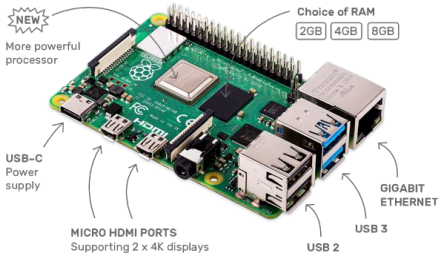
Deep learning requires more resources during the training phase than during the deployment phase

Training

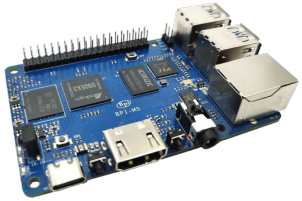
Running

Platforms for edge computing

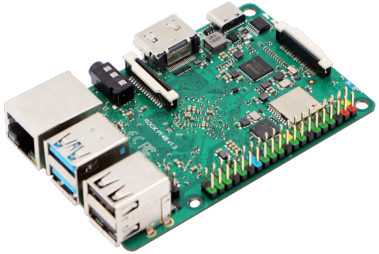
Small computers



Raspberry Pi



Banana Pi



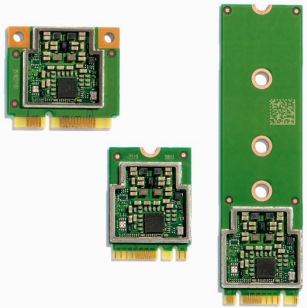
Rock Pi

Platforms for edge computing

Deep learning-specific processors

Google Coral edge TPU

Carrier board



PCIe modules

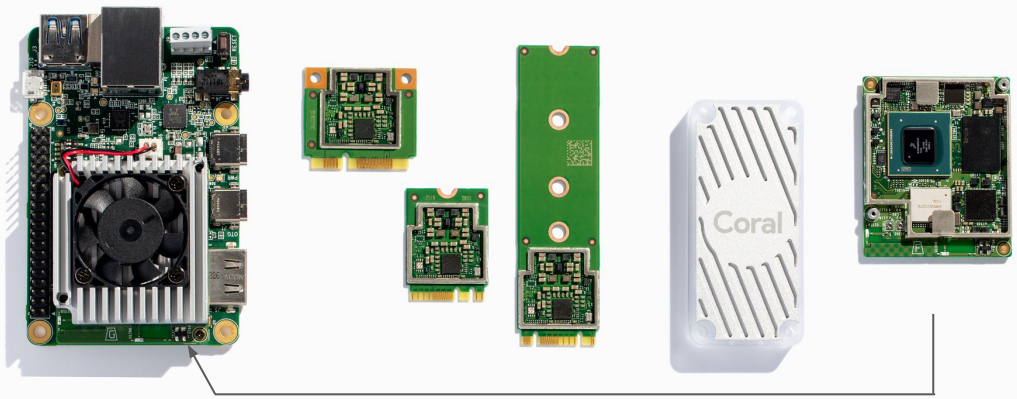
USB accelerator

System on Module

Solderable module

Platforms for edge computing

Deep learning-specific processors



Carrier board

Google Coral edge TPU

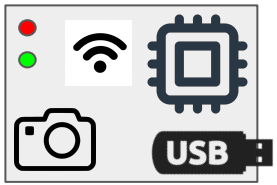
Solderable module

Platforms for edge computing

Carrier boards (development, evaluation boards)

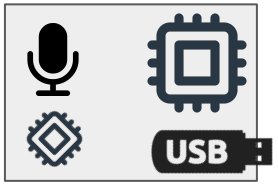
Device

Carrier board



More general

User's application



More specific
(Power budget, environmental conditions, additional hardware)

Platforms for edge computing

Deep learning-specific processors

Jetson Nano

Jetson Tx2

Jetson Xavier

Nvidia Jetson

Jetson Nano Dev board

Platforms for edge computing

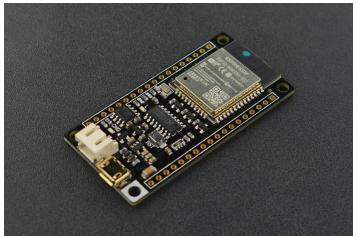
Microcontrollers

- A number of basic components you would find in a computer (processors, memory, oscillators, etc) in one chip
- Lower power consumption and computing power

Platforms for edge computing

Microcontrollers

DFRobot Firebeetle



With Espressif's esp32

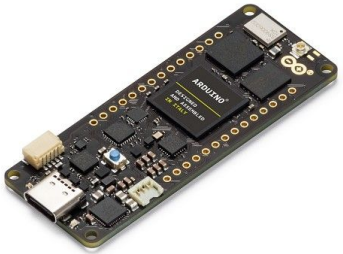
Arduino nano ble sense

With Nordic's nRF52840

NUCLEO-F746ZG

With ST's STM32F746

Arduino Portenta H7



With ST's STM32H747

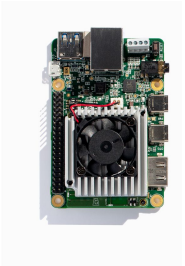
Platforms for edge computing

Software

TensorFlow → TensorFlow Lite → TensorFlow Lite For microcontrollers

train/run

run



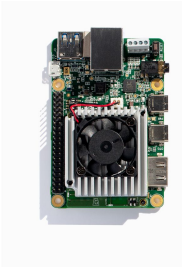
run

Platforms for edge computing

Software

train/run

run



run

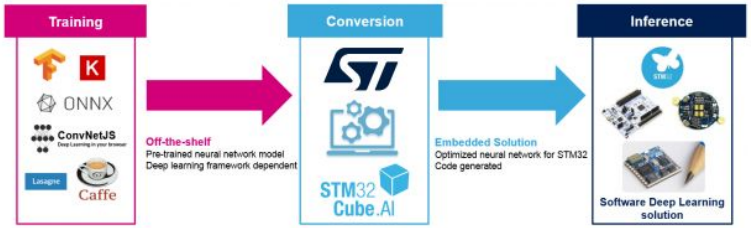
Platforms for edge computing

Software

deepC
(github.com/ai-techsystems/deepC)

Platforms for edge computing

Software

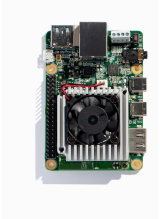


(https://www.st.com/en/embedded-software/x-cube-ai.html)

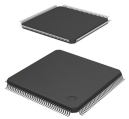
Constraints

Knowledge of electronics/
low-level programming

Small computers/ DL SoMs



Microcontrollers

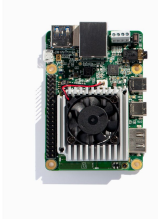


Power consumption
Computing power
Storage capacity
Memory

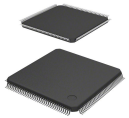
Constraints

Knowledge of electronics/
low-level programming

Small computers/ DL SoMs



Microcontrollers



3~10W
4x 1.5GHz
10s, 100s GBs
1~8GB

Power consumption
Computing power
Storage capacity
Memory

100s mW~2W
40~500 MHz
256KB ~ 1MB
100s KB, few MB

Conclusions

- The kind of model you can run in the field will depend on what devices your system can afford
- Small computers and DL SoMs are capable of running relatively complex models
- Microcontrollers will run simpler models, which might be enough for some applications

Thank You!