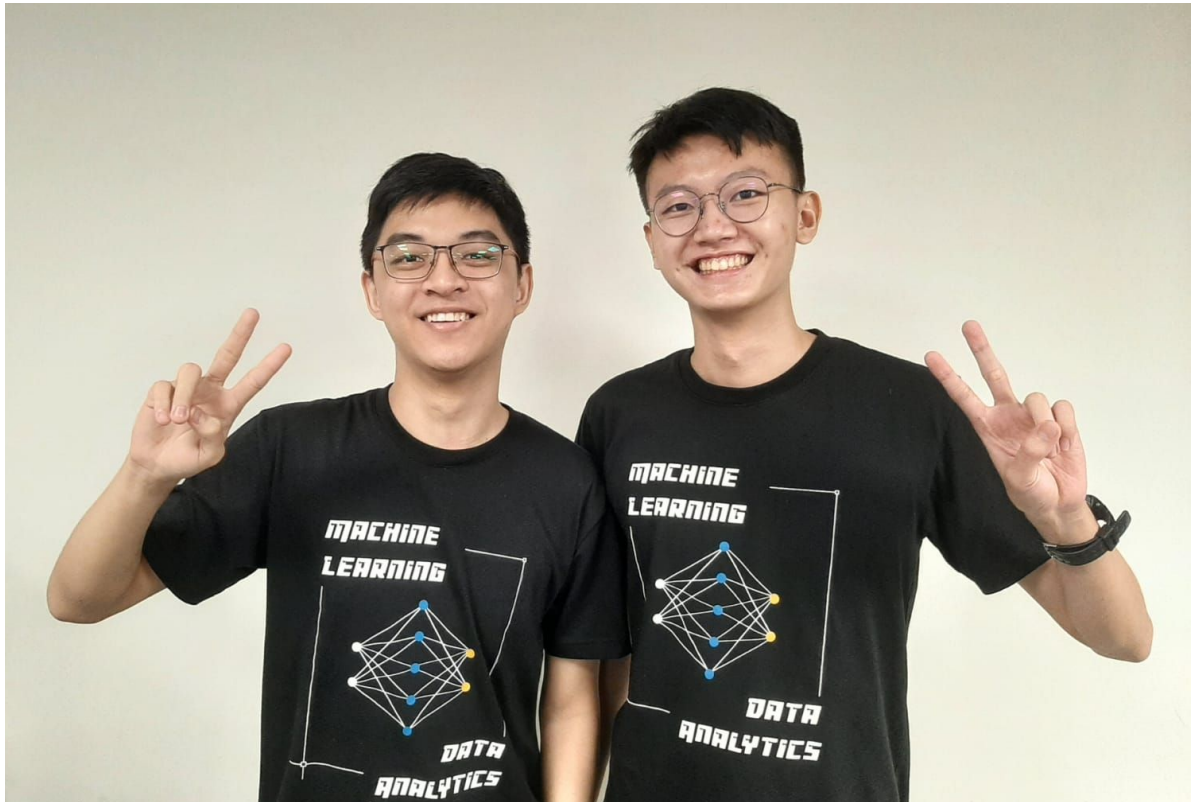
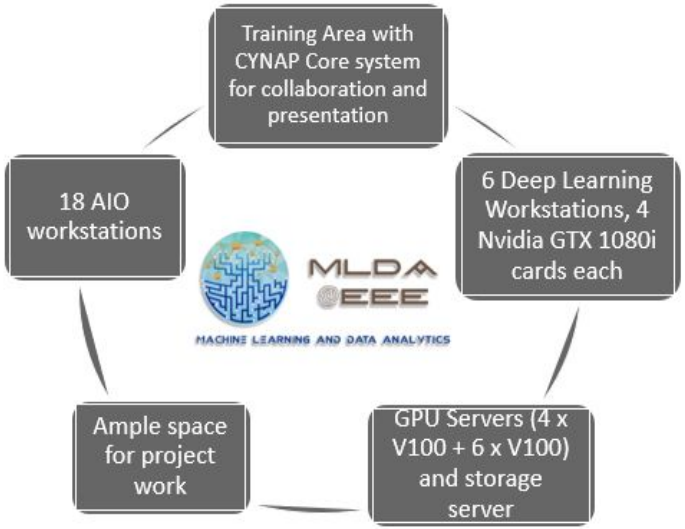
Edge ML: From Cloud to Your Fingertips

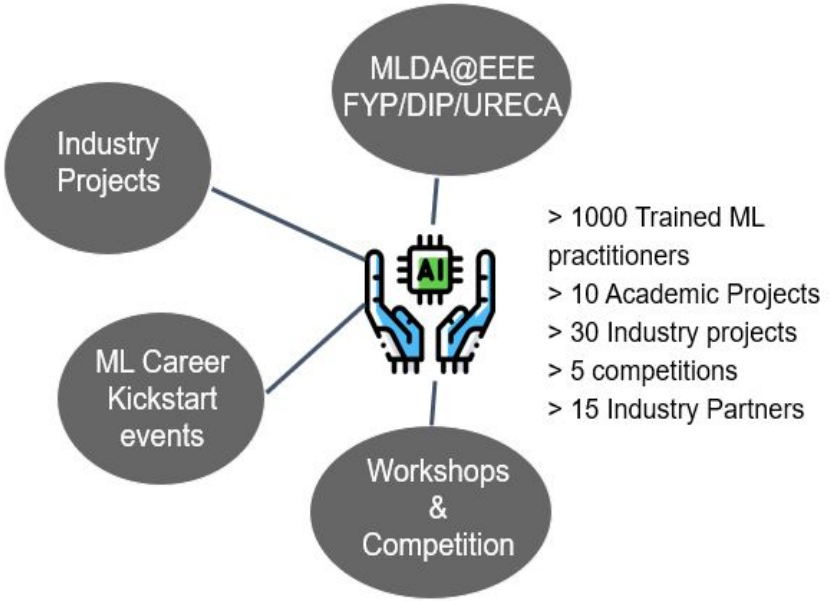
Instructors: Philip, Yong Hao


MACHINE
LEARNING
DATA
ANALYTICS
MACHINE
LEARNING
DATA
ANALYTICS

MLDA's Mission

Provide an integrated platform for EEE/IEM students to learn and implement Machine Learning, Data Science & AI, as well as facilitate connections with the industry.





Edge ML: New possibilities through machine learning directly on microcontrollers

Workshop Outline

Part 1

1. What and Why?
2. Special Techniques
3. Recent Developments

Part 2

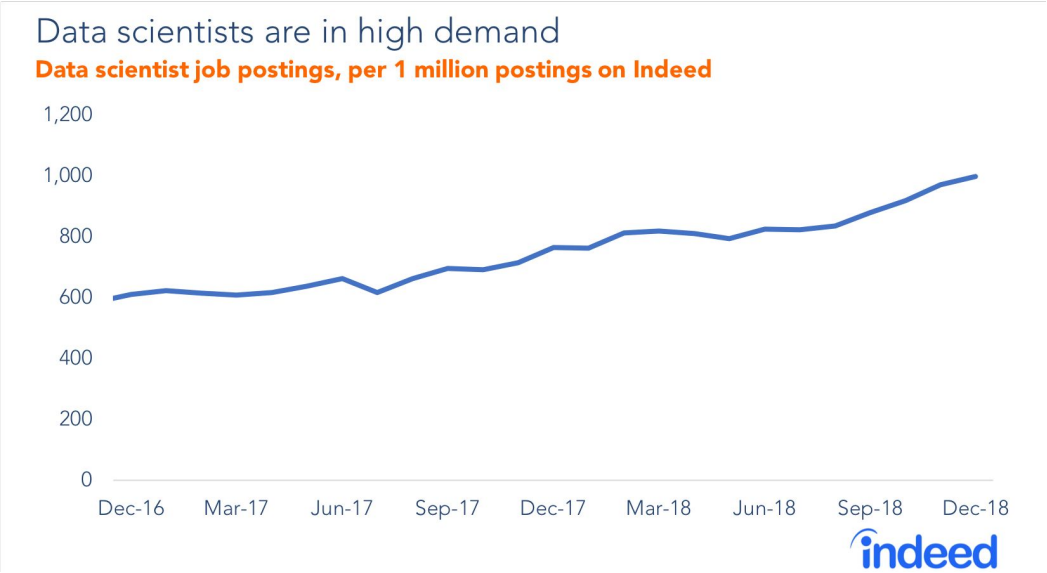
1. Demo Session with ESP32
2. Google Colab Demo
3. Hands on with Edge Impulse

Edge ML
What and Why?

We are shifting to a data-centric world

Estimates: We generate 145 trillion MB per day!

Many companies use data to improve their business / product / service.



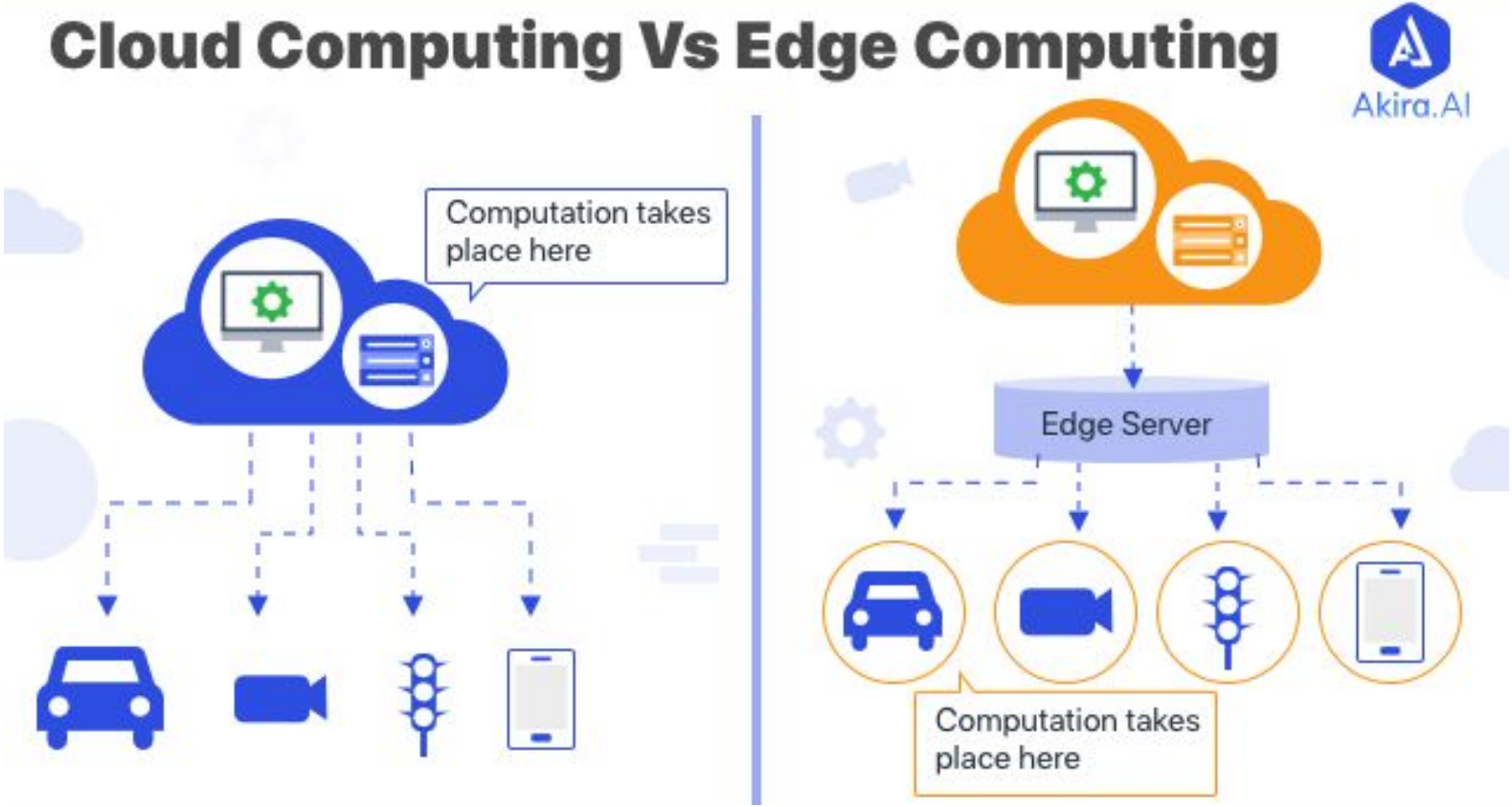
We are shifting to a data-centric world

Estimates: We generate 145 trillion MB per day!

Many companies use data to improve their business / product / service.

Top 10 Best Jobs in America in 2020

Rank	Job Title	Median Base Salary	Job Satisfaction	Job Openings
1	Front End Engineer	$105,240	3.9	13,122
2	Java Developer	$83,589	3.9	16,136
3	Data Scientist	$107,801	4.0	6,542
4	Product Manager	$117,713	3.8	12,173
5	Devops Engineer	$107,310	3.9	6,603
6	Data Engineer	$102,472	3.9	6,941
7	Software Engineer	$105,563	3.6	50,438
8	Speech Language Pathologist	$71,867	3.8	29,167
9	Strategy Manager	$133,067	4.3	3,515
10	Business Development Manager	$78,480	4.0	6,560


Cloud Computing Vs Edge Computing
Akira.AI
Computation takes place here
Edge Server
Computation takes place here

Edge Devices

and
many
more
...

Rule-based → Probabilistic-based

Application of Microcontroller in Day to Day Life Devices:

- Light sensing & controlling devices.
- Temperature sensing and controlling devices.

Microcontroller with ML / AIoT

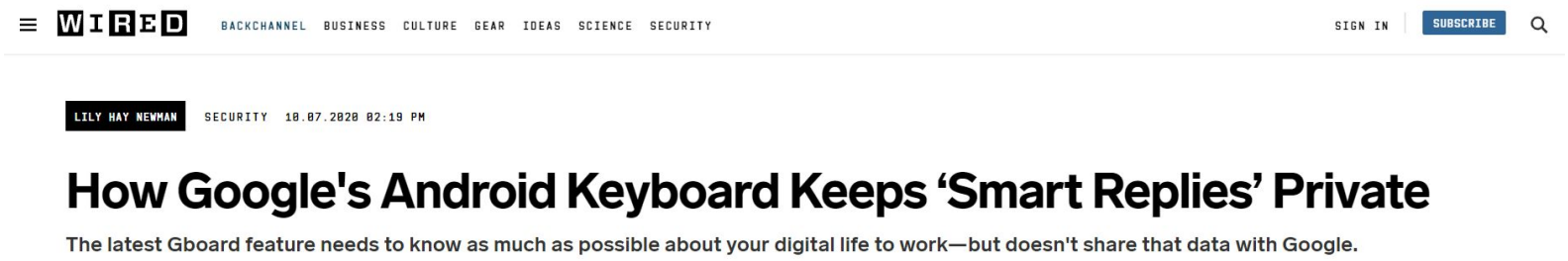
- Sensor fusion to report health statistics / emergency alerts (connectivity not needed)
- Pattern / object detection of objects on conveyor belt

Benefits of Edge Computing / AI

- Support for low-latency use cases and fast response times
- Resilience against connectivity issues - Many areas have slow / no internet connection
- Lower costs:
    - Upfront: Embedded devices
    - Long-term: Network Bandwidth, Cloud, Energy

Benefits of Edge Computing / AI

- Data Privacy and Security
  - Personalised data kept locally only - Eg. GBoard



FEDERATED LEARNING FOR MOBILE KEYBOARD PREDICTION

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays
Sean Augenstein, Hubert Eichner, Chloé Kiddon, Daniel Ramage

Google LLC,
Mountain View, CA, U.S.A.
{harda, kanishkarao, mathews, swaroopram, fsb
saugenst, huberte, loeki, dramage}@google.com

ABSTRACT

We train a recurrent neural network language model using a distributed, on-device learning framework called federated learning for the purpose of next-word prediction in a virtual keyboard for smartphones. Server-based training using stochastic gradient descent is compared with training on client devices using the `FederatedAveraging` algorithm. The federated algorithm, which enables training on a higher-quality dataset for this use case, is shown to achieve better prediction recall. This work demonstrates the feasibility and benefit of training language models on client devices without exporting sensitive user data to servers. The federated learning environment gives users greater control over the use of their data and simplifies the task of incorporating privacy by default with distributed training and aggregation across a population of client devices.

***Index Terms*—** Federated learning, keyboard, language modeling, NLP, CIFG.

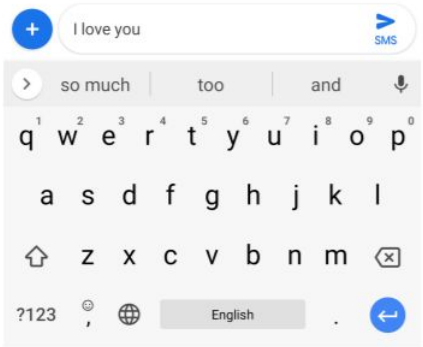


Fig. 1. Next word predictions in Gboard. Based on the context "I love you", the keyboard predicts "and", "too", and "so much".

https://arxiv.org/abs/1811.03604

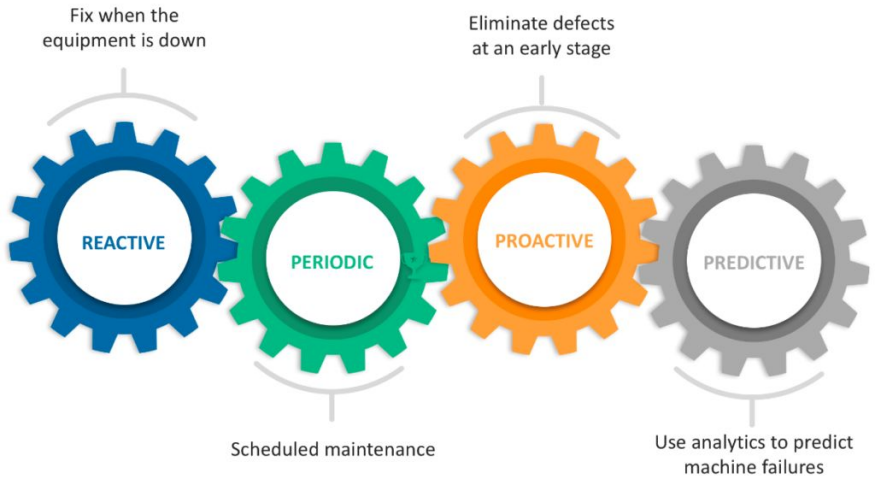
The CIFG architecture is advantageous for the mobile device environment because the number of computations and the parameter set size are reduced with no impact on model performance. The model is trained using TensorFlow [22] without peephole connections. On-device inference is supported by TensorFlow Lite[2].

Tied input embedding and output projection matrices are used to reduce the model size and accelerate training [23, 24].

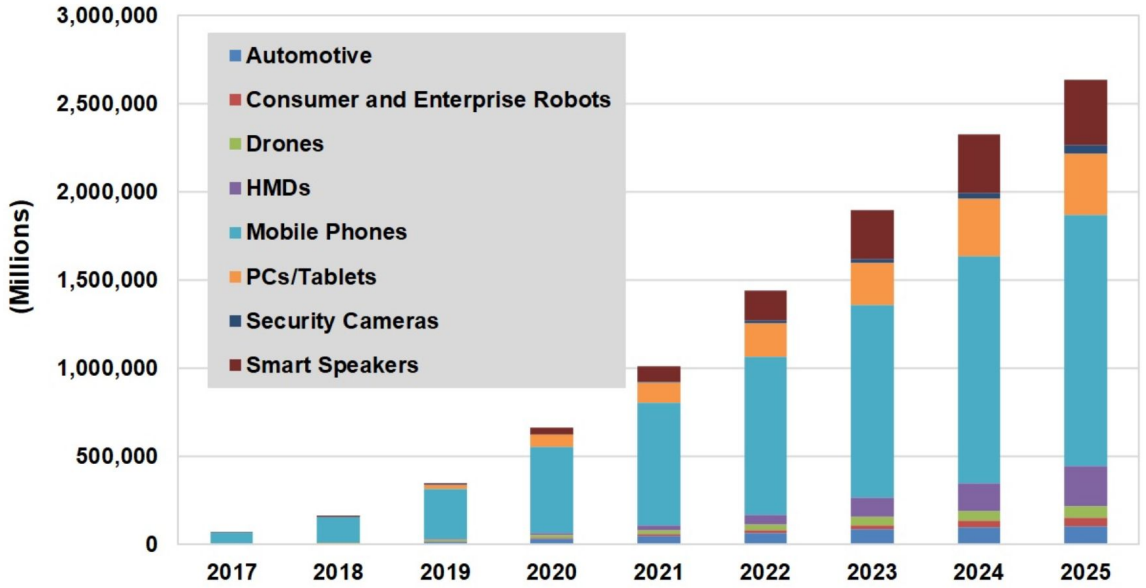
Industry 4.0 Shifts

1) NVIDIA introducing on-device GPU in security cameras for local detection.
   → Jetson TX2 embedded AI supercomputer

2) Manufacturing industry: Predictive maintenance



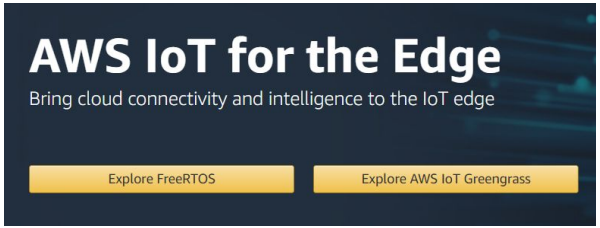
AI Chip Architectures Race To The Edge




arm
life.augmented
hackster.io
AN AVNET COMMUNITY
TensorFlow
THE THINGS NETWORK
TINY ML
Eta Compute
ARDUINO
MICROCHIP
edge ai + vision ALLIANCE
Himax
seeed
NORDIC SEMICONDUCTOR
SILICON LABS
SONY
NVIDIA
OpenMV
Texas Instruments

Edge AI will not completely replace Cloud!

- Edge will work in complement with the Cloud
- User data can be ‘sanitised’ before uploading to Cloud
  - Eg. Data collection, Federated learning
- High demand computational tasks will still need Cloud
- Cloud based entities are investing in Edge too
  - Complement service



Edge ML
Special Techniques

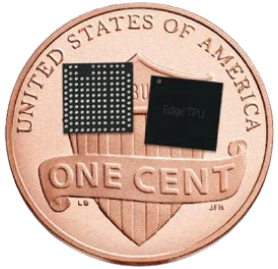
What are the challenges? Why only now?

Low compute resources

No parallelism during model inference (pure CPU based)

Moore's law

Specialised chips - Google's Edge TPU, AI SOCs



Remember to check back model performance!

If optimization steps causes model performance to suffer, then investigate …

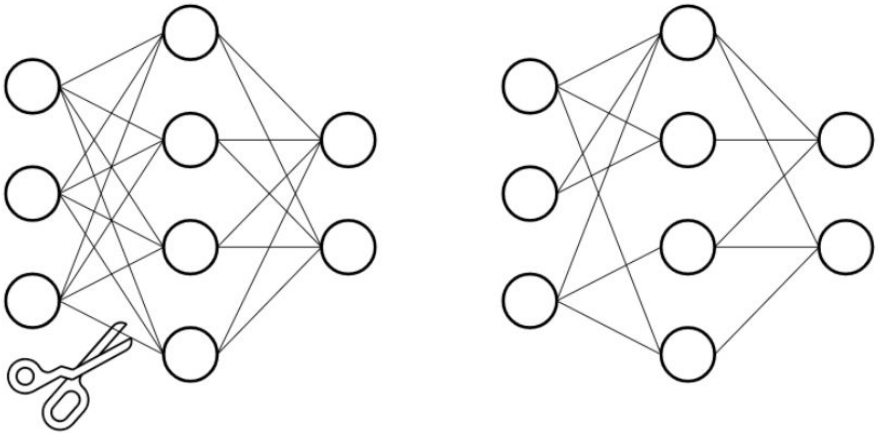
Further reading:

https://www.oreilly.com/content/compressing-and-regularizing-deep-neural-networks/

https://blog.dataiku.com/making-neural-networks-smaller-for-better-deployment-solving-the-size-problem-of-cnns-using-network-pruning-with-keras

Technique 1 - Weight Pruning

Reduce the number of parameters and operations involved in the computation by removing connections, and thus parameters, in between neural network layers.

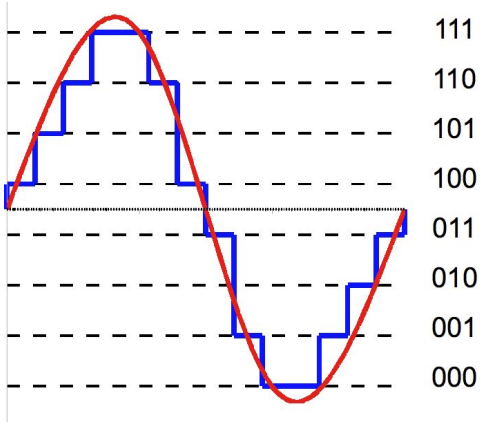


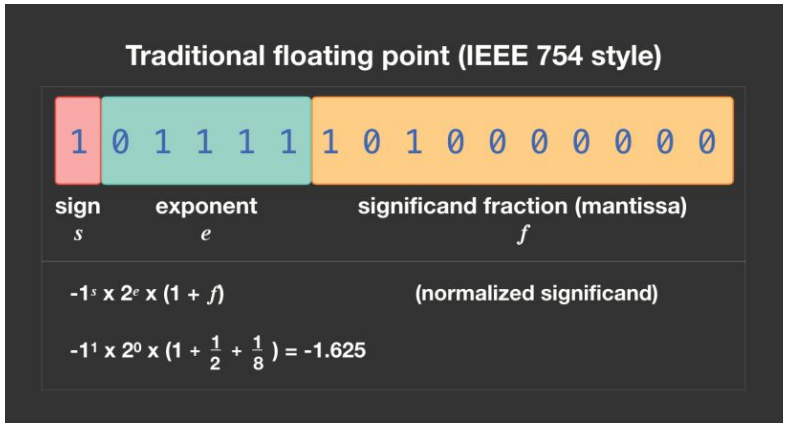
Before pruning

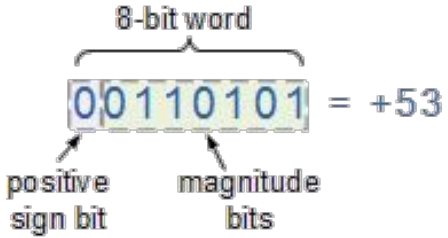
After pruning

Technique 2 - Quantization

Integer quantization is a general technique that reduces the numerical precision of the weights and activations of models to reduce memory and improve latency.

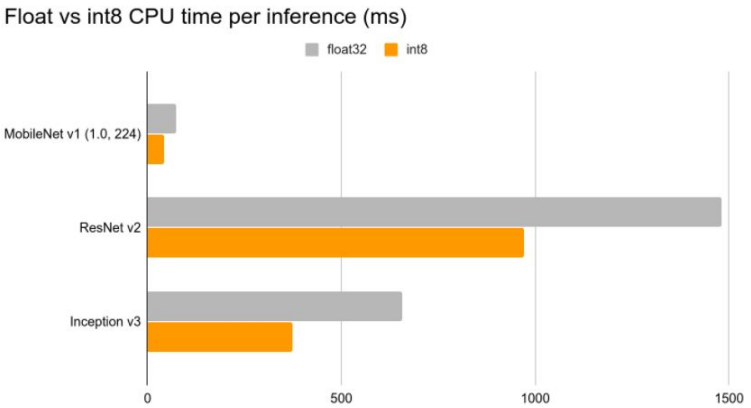






Technique 2 - Quantization

Int is more efficient!



Model	Float baseline	Quantization during training	Quantization after training
MobileNet v1 (1.0, 224)	70.95%	69.97%	69.568%
ResNet v2	76.8%	76.7%	76.652%
Inception v3	77.9%	77.5%	77.782%

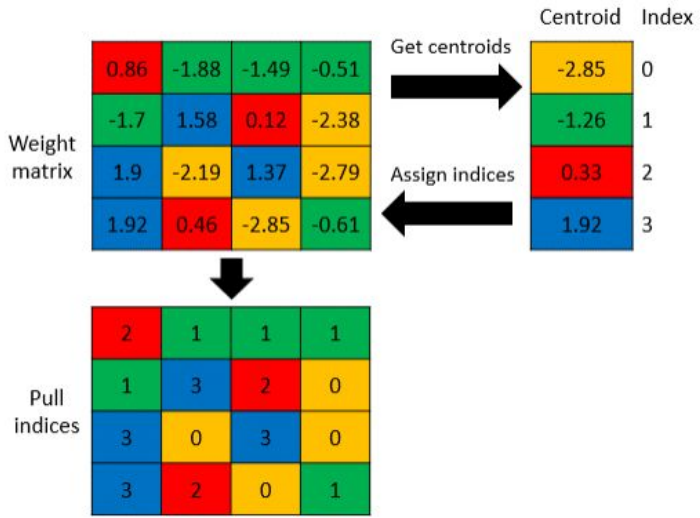
Technique 3 - Network Pruning: Rank-Prune-Retrain

Applied as a general technique not just for Edge ML

- Reduced complexity of model
- Reduced number of parameters

Technique 4 - Weight Clustering

Reduces the size of your model by replacing similar weights in a layer with the same value



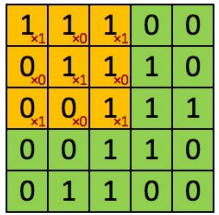
Specially-designed Model Architectures

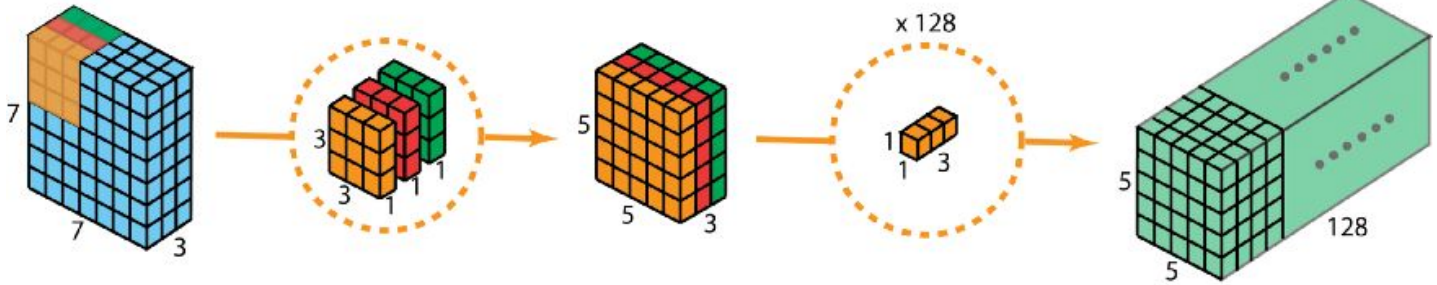
MobileNet: Depth-wise Separable Convolutions

https://arxiv.org/abs/1704.04861

Table 4. Depthwise Separable vs Full Convolution MobileNet

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
Conv MobileNet	71.7%	4866	29.3
MobileNet	70.6%	569	4.2

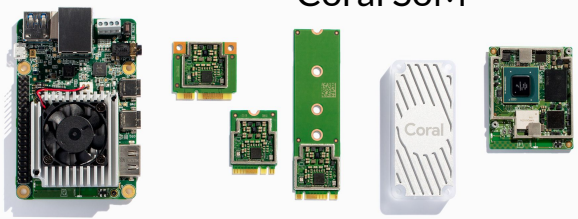




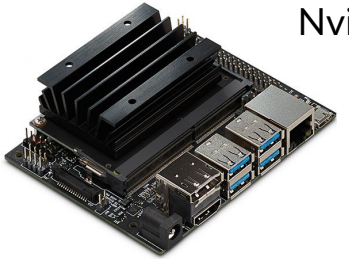
Edge ML
Deployment - Steps and Frameworks

Edge Devices

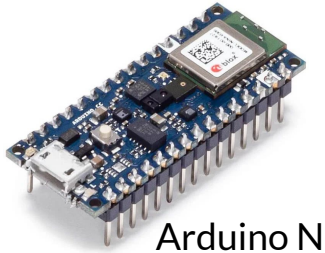
Coral SoM



Nvidia Jetson Nano



AI Hardware Accelerators

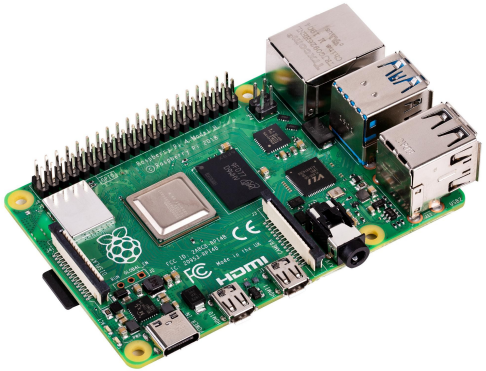


Arduino Nano 33 BLE Sense

Seeeduino XIAO

Microcontrollers

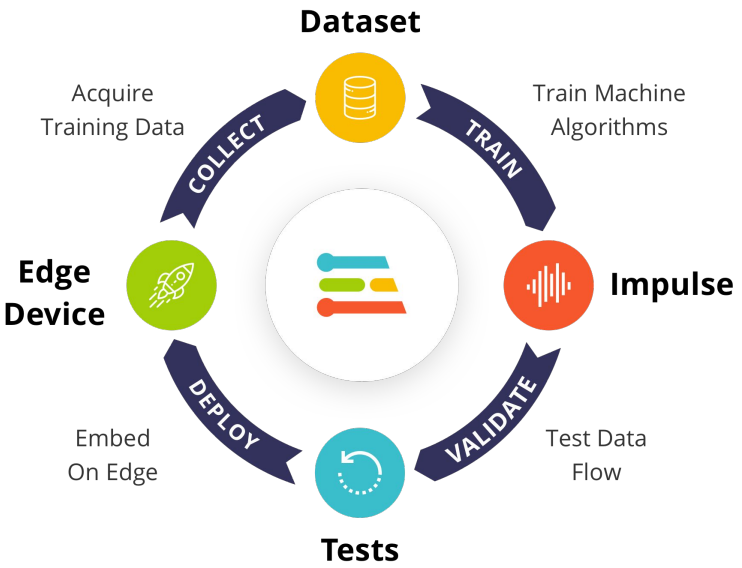
Raspberry Pi 4



Micro-computer

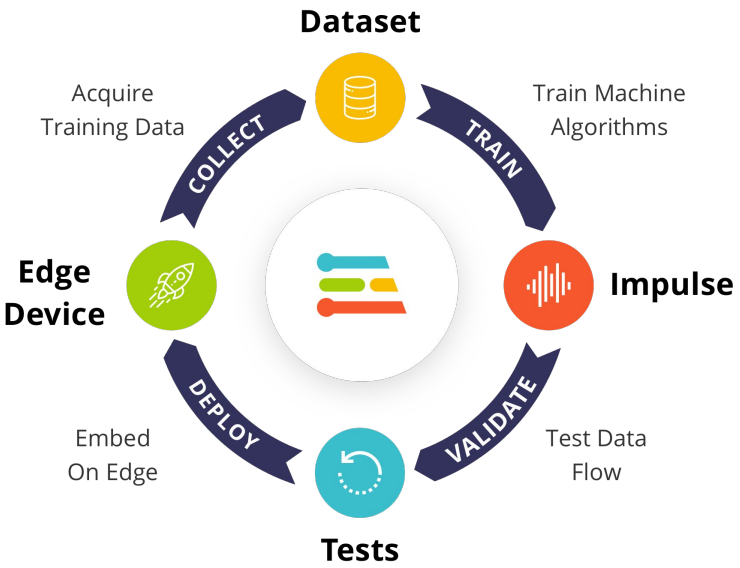
What are the common steps to deploy edge ML?

1. Collecting, exploring, and evaluating dataset
2. Feature engineering (digital signal processing) -> Heavylifting
3. Training and evaluating machine learning model



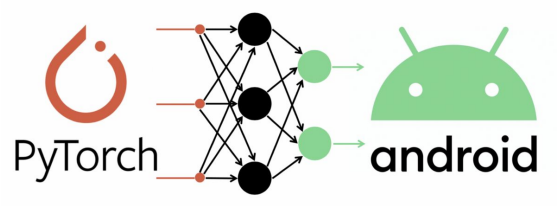
What are the common steps to deploy edge ML?

4. Size and efficiency optimization of model (most commonly quantization)
5. Deployment (commonly C++ library) with required operator kernels
6. Writing and tuning an application that interprets the model's output and uses it to make decisions.



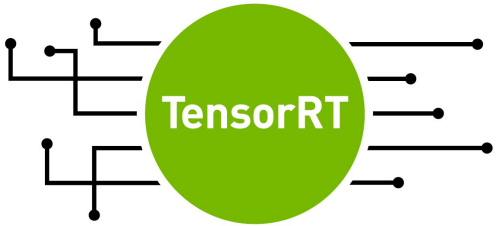
Common frameworks for TinyML models

- TensorFlow Lite
- CoreML (Apple library)
- PyTorch Mobile (Facebook's Pytorch Deep Learning Library)

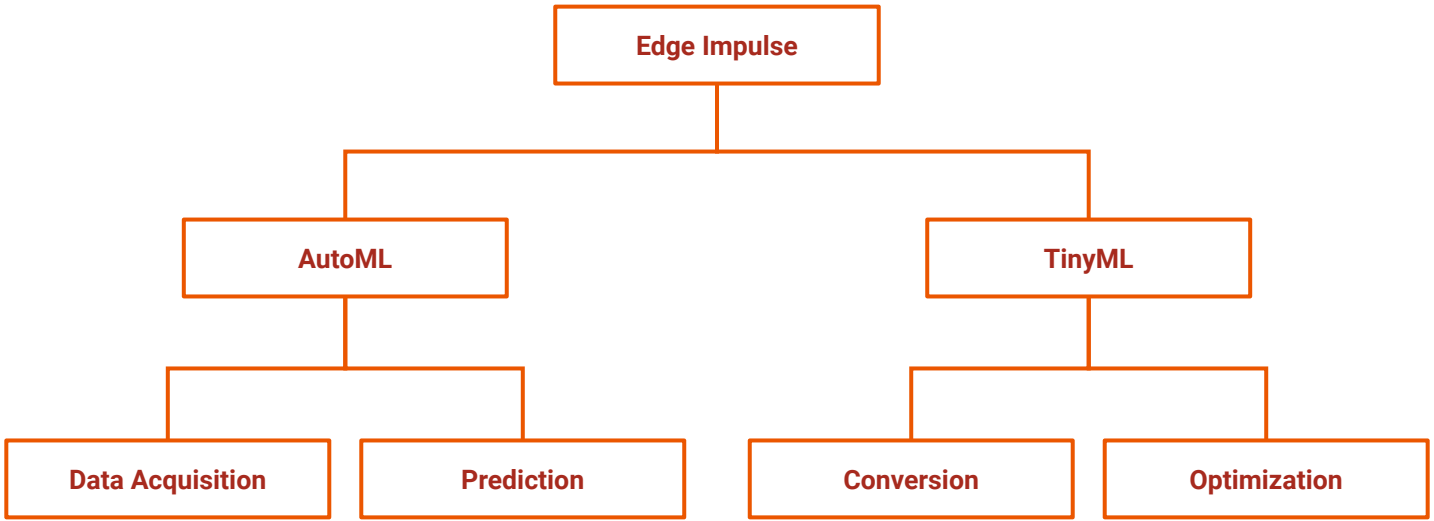


Other Edge ML Technologies

- ONNX
- TensorRT (Nvidia)
- Openvino (Intel)
- Tensorflow.js (Javascript)



EDGE IMPULSE



Edge Impulse

Advanced DSP blocks

- Generates important features from raw sensor data
- Reduce feature engineering workload for neural network model

Latency and Memory Usage Estimate

- Comparison can be done before deployment and changes can be made immediately

Direct compilation to C++ Source Code

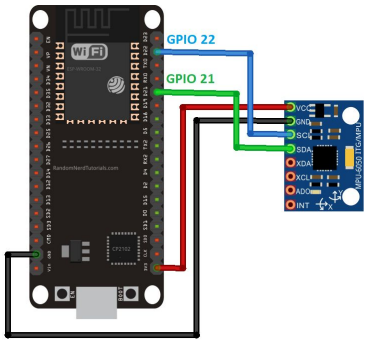
- No interpreter needed
- Shifts data easily into ROM
- Linker knows exactly which operations are being used

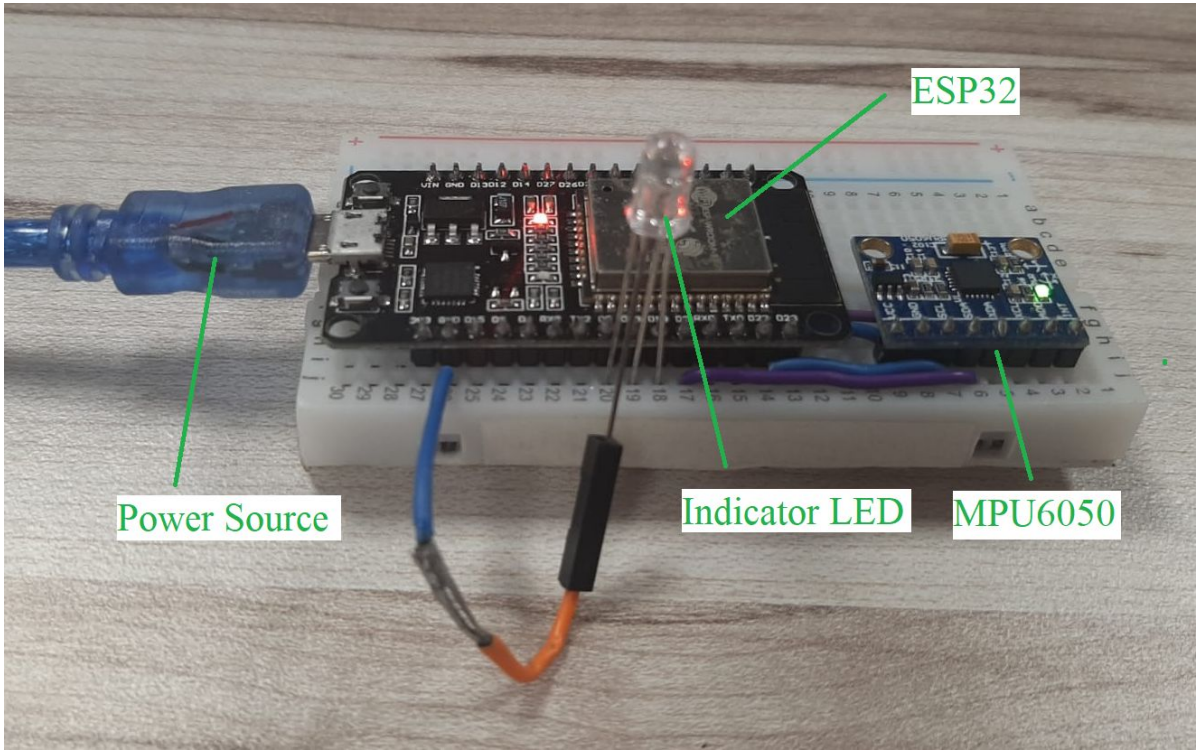
Break Time!

Please make an Edge Impulse account (on PC, not mobile):
https://www.edgeimpulse.com/

Edge ML
Demo Session with ESP32

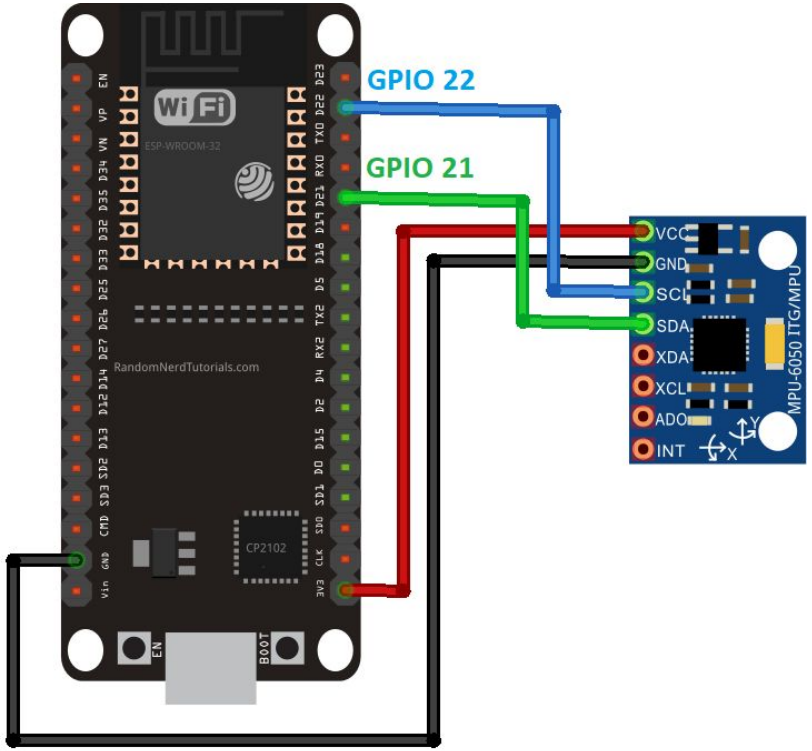
Gesture Recognition on ESP32





Pin Diagram

https://www.hackster.io/Yukio/gesture-classification-with-esp32-and-tinyml-dab252



Edge ML
Google Colab Demo

Flower Classification with TFLite Model Maker

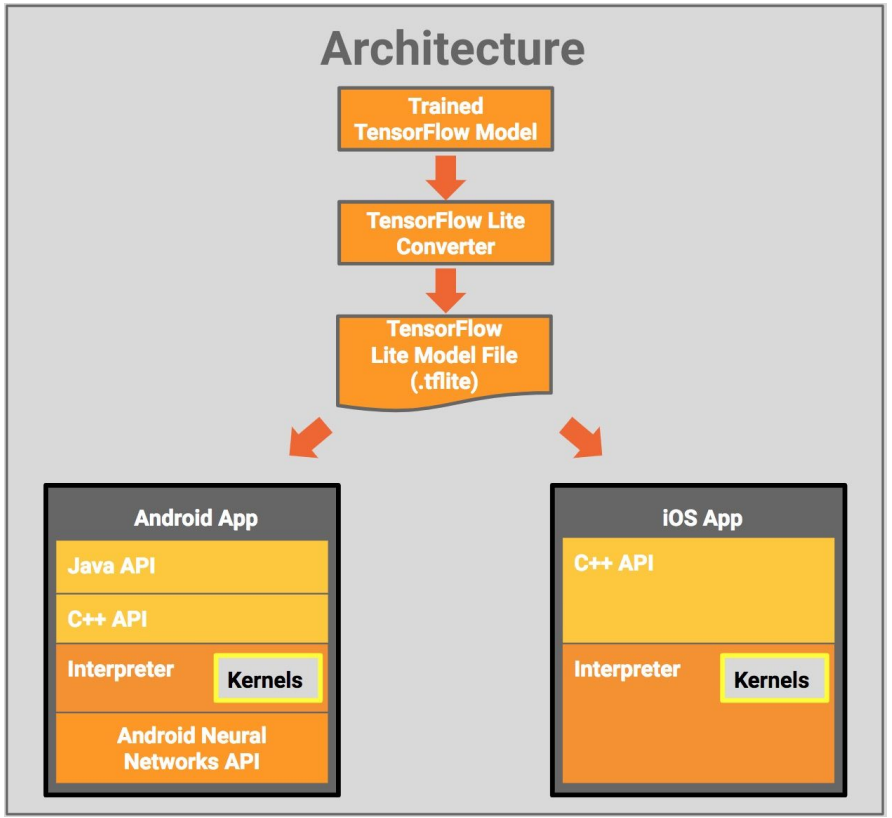
https://colab.research.google.com/drive/1iXDz6r9kLeQ6xjJ1O0F-iV8LZ9A_MV8x

Optional

Tflite Usage

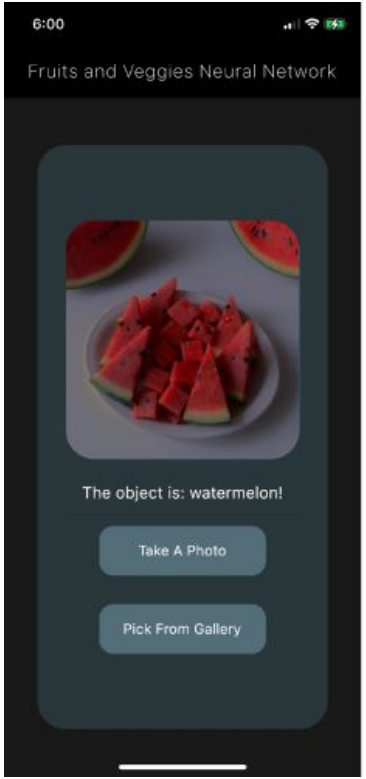
TensorFlow Lite API

- Python
- Android (Java)
- Swift
- C++



Sample App

https://medium.com/google-cloud/on-device-machine-learning-train-and-run-tensorflow-lite-models-in-your-flutter-apps-15ea796e5ad4



Edge ML
Hands on with Edge Impulse

Keyword Detection on Mobile Phone

What you need:

- 1 x Computer
- 1 x Mobile Phone
- Internet connection

4 categories: Yes, No, Noise, MLDA

https://docs.edgeimpulse.com/docs/keyword-spotting

Photo Time!

Upcoming Workshops!

We value your Feedback!

https://bit.ly/dlwEdge