AI for Edge with Keras

Agenda 🚀

What is EdgeAI?

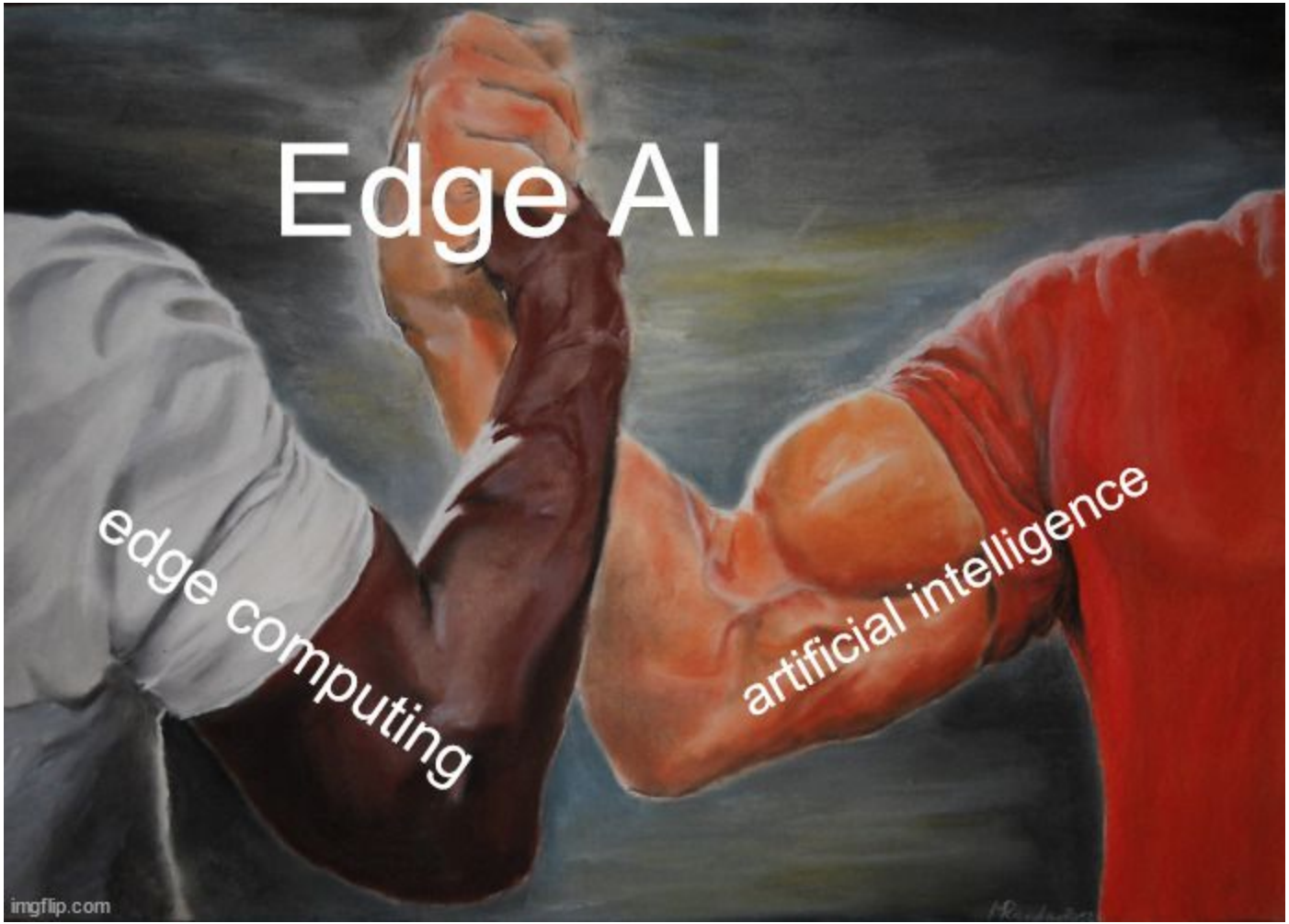
Cloud to Edge

Key Metrics

Model Optimization Techniques

Case Studies

Q & A


Edge AI
edge computing
artificial intelligence
imgflip.com

What is EdgeAI?

AI on the Edge refers to the practice of running artificial intelligence (AI) algorithms directly on endpoint devices, such as IoT devices, smartphones, drones, or sensors, rather than in a centralized data center or cloud environment.

The **Edge** refers to the computational capabilities embedded in these devices at the "edge" of the network, as opposed to centralized servers.

The benefits of On-Device Machine Learning

Low Latency

Unlock new user experiences by processing text, audio and video in real-time

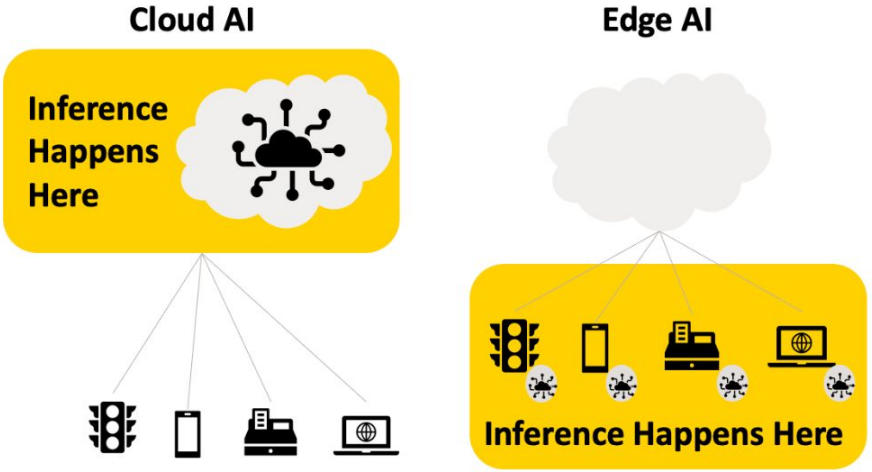
Keep data on-device

Perform inference locally without sending user data to the cloud

Works offline

No need for a network connection or running a service in the cloud

Cloud to Edge



	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Cloud compute cost (USD)
Transformer (65M parameters)	Jun, 2017	27	26	$41-$140
Transformer (213M parameters)	Jun, 2017	201	192	$289-$981
ELMo	Feb, 2018	275	262	$433-$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	$3,751-$12,571
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	$942,973-$3,201,722
GPT-2	Feb, 2019	-	-	$12,902-$43,008

Note: Because of a lack of power draw data on GPT-2's training hardware, the researchers weren't able to calculate its carbon footprint.

Training a single AI model can emit as much carbon as five cars in their lifetimes

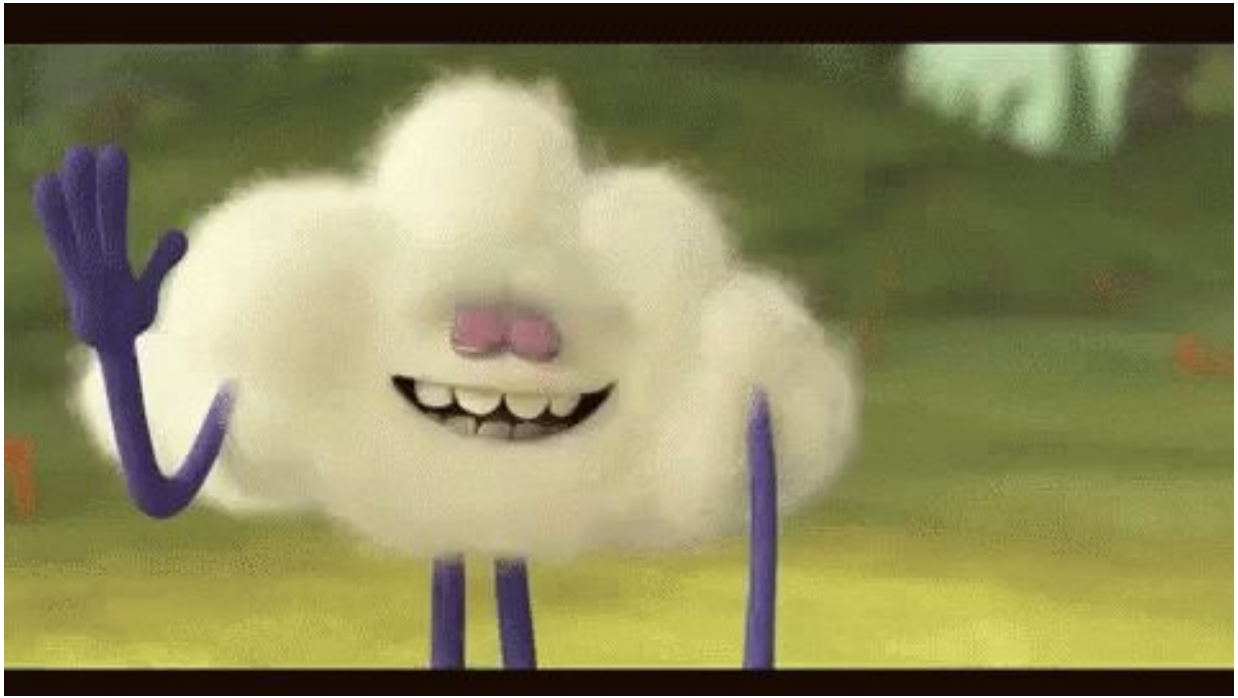
A Simple Keras Model

```
from tensorflow import keras
```

```
model = keras.Sequential([
    keras.layers.Dense(units=16, activation='relu', input_shape=(2,)),
    keras.layers.Dropout(rate=0.2),
    keras.layers.Dense(units=8, activation='relu'),
    keras.layers.Dropout(rate=0.2),
    keras.layers.Dense(units=3, activation='softmax')
])
```

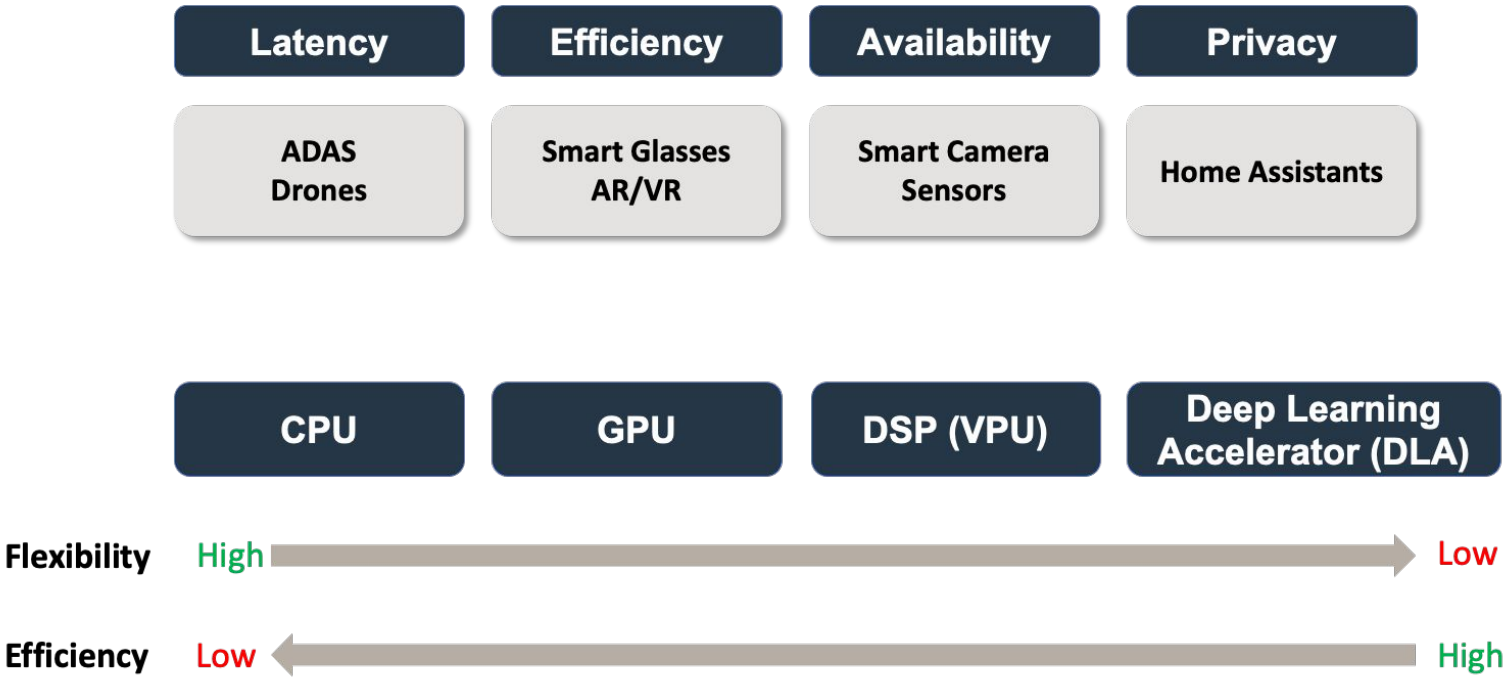
Train it! Deploy it!

```
model.save('my_model.h5')
```



Source: Tenor

Challenges in EdgeAI Deployment



Understanding Key Metrics

- Size Reduction
- Accuracy
- Throughput
- Latency
- Memory Bandwidth
- Power & Energy Consumption

Model Optimization Techniques

“Model optimization refers to the process of improving the performance, efficiency, and resource utilization of a trained machine learning model. This involves various techniques aimed at achieving better results while using fewer computational resources.”

We’ll talk about

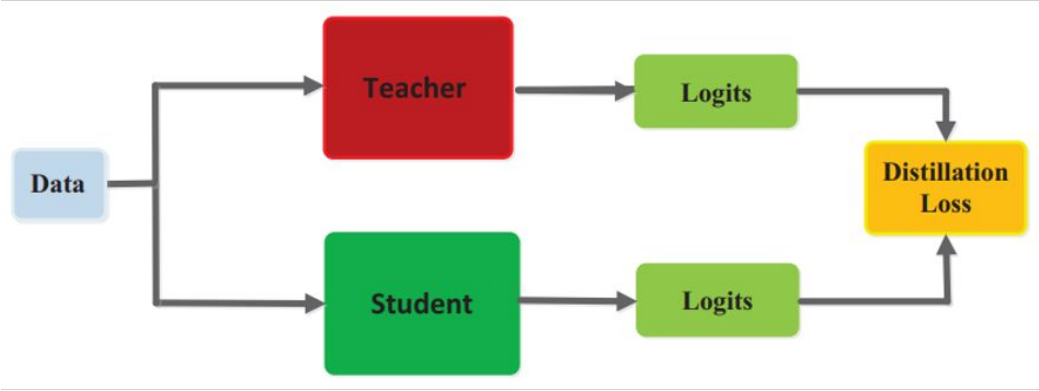
- Knowledge Distillation
- Quantization

Optimize models to reduce size, latency and power for negligible loss in accuracy

Source: TensorFlow Blog

Knowledge Distillation

"A technique where a smaller model (student) is trained to reproduce the behavior of a larger model (teacher) or an ensemble of models, often leading to a compact model with comparable performance"



Source: Knowledge Distillation: A Survey

Knowledge Distillation

Size Reduction
Up to 10% – 50%
w.r.t teacher model

Throughput
2x – 10x higher
w.r.t teacher model

Memory Bandwidth
50% – 75% reduction,
depending on student model
design

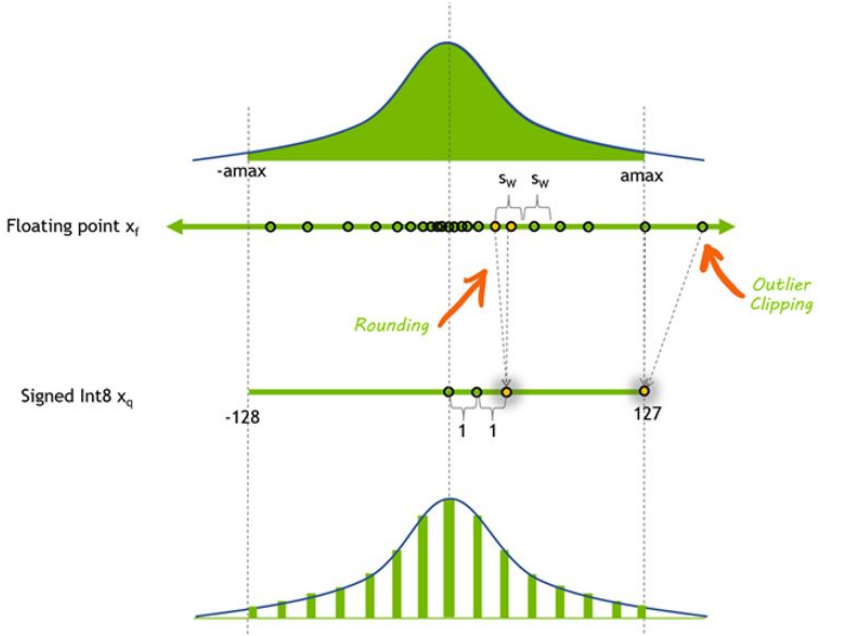
Accuracy
1% - 5% drop,
w.r.t teacher model

Latency
2x – 10x reduction

Power& Energy Efficiency
2x – 10x lower consumption

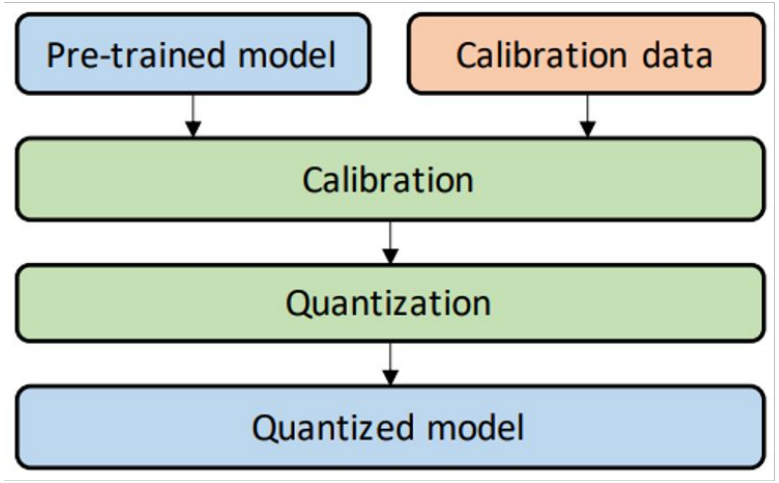
Quantization

"The process of reducing the numerical precision of model parameters by mapping it from a large number of possible values to a reduced set of values"

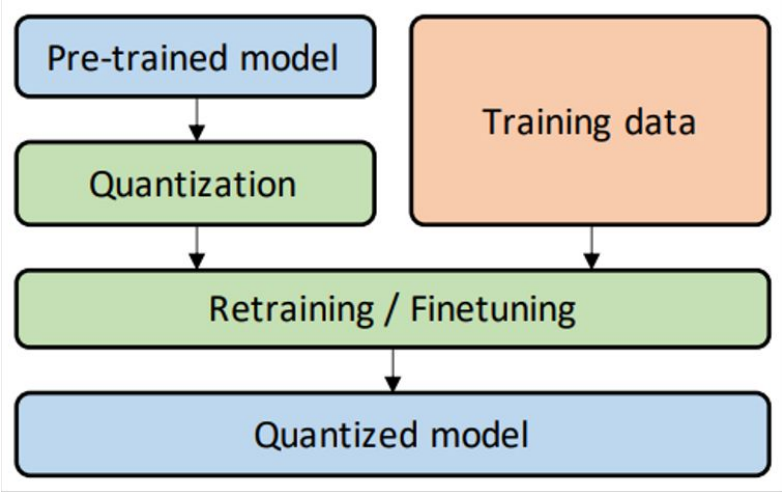


Achieving FP32 Accuracy for INT8 Inference Using Quantization Aware Training with NVIDIA TensorRT

Quantization



Post Training Quantization



Quantization Aware Training

A Survey of Quantization Methods for Efficient Neural Network Inference

Quantization

Size Reduction
Up to 50% – 75%
w.r.t FP32 model

Throughput
2x – 4x higher

Memory Bandwidth
50% – 75% reduction,
depending on bit-width

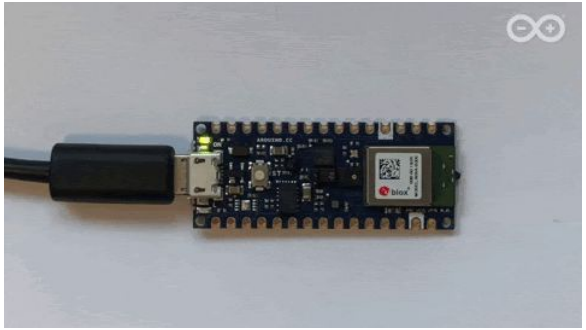
Accuracy
1% - 5% drop,
depending on bit-width and
quantization technique

Latency
2x – 3x reduction

Power& Energy Efficiency
2x – 3x lower consumption

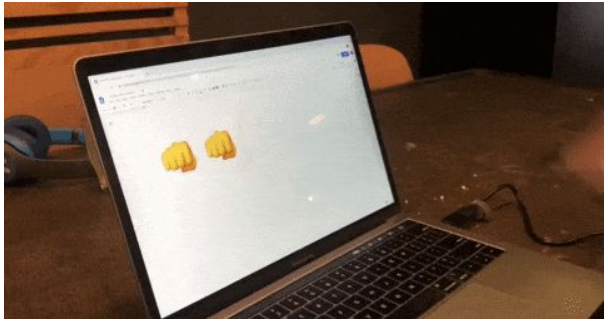
Few of us after looking at Model Optimization techniques:

Case Study 1: AI on embedded



Voice Activation

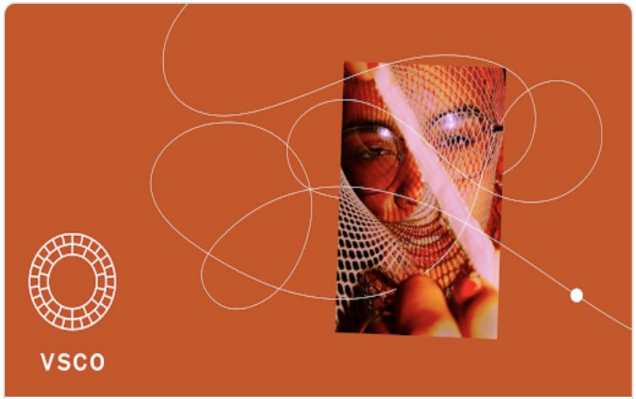
Object detection



Gesture Recognition

Source: arduino.cc
Raspberry Pie

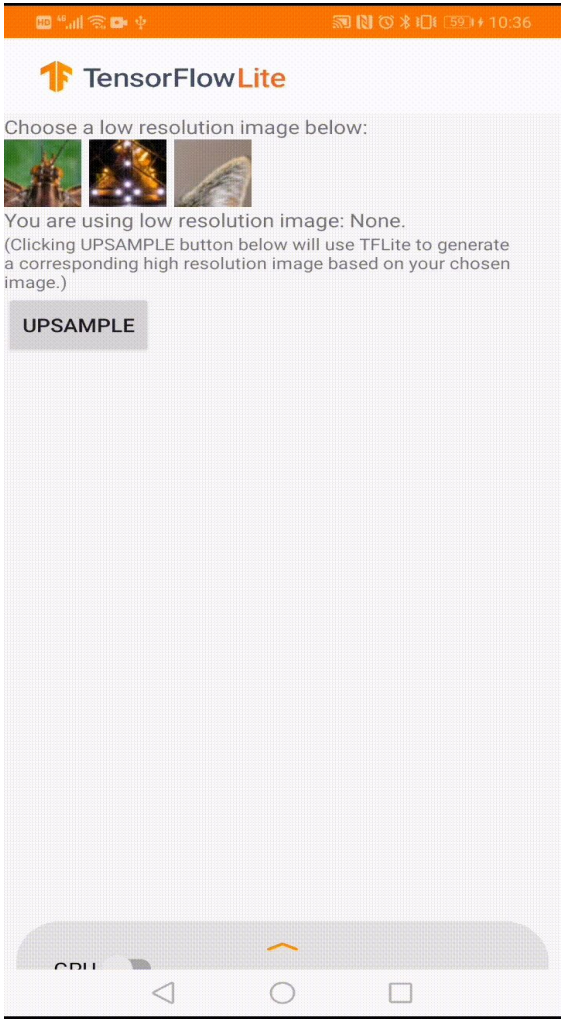
Case Study 2: On Mobile



TENSORFLOW LITE

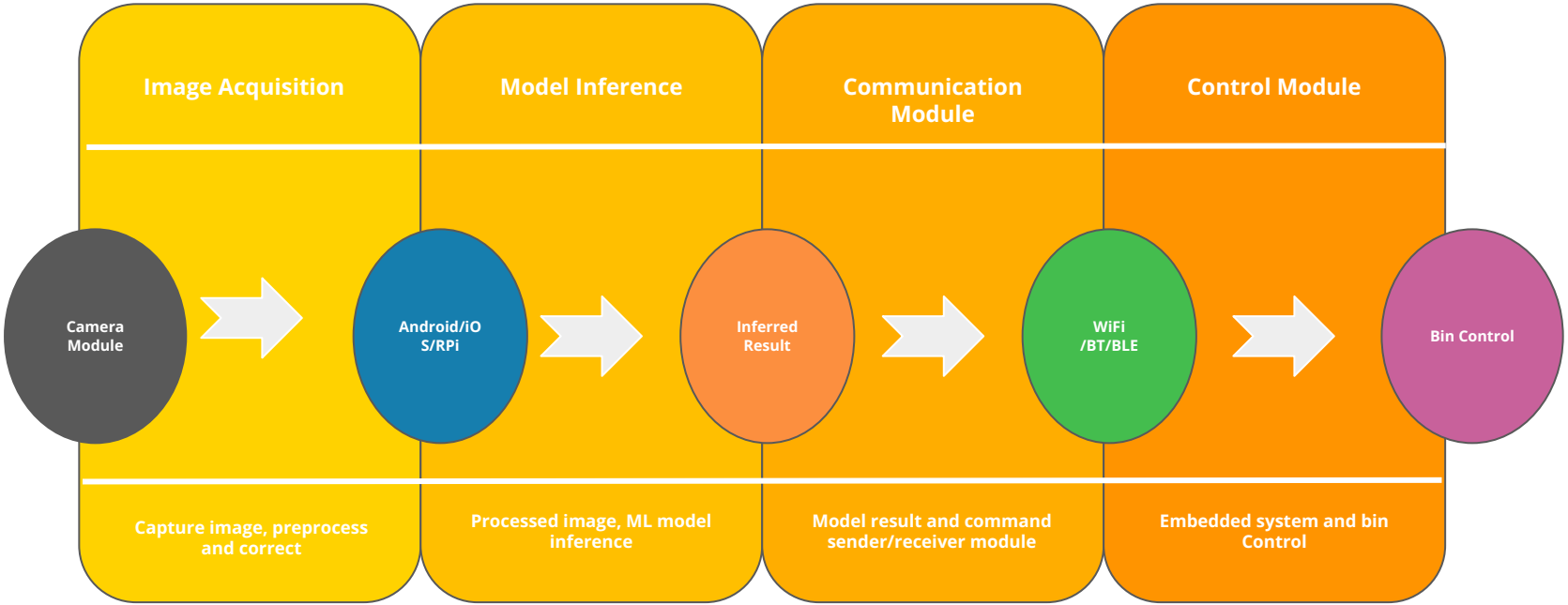
VSCO uses on-device ML to recommend image presets

With hundreds of VSCO photo presets to choose from, helping users find and try new presets was a challenge. But with on device ML, the VSCO app now understands uploaded images and suggests presets that best complement them via the "For this photo" feature.



Source: TF.org

Let’s club the two: Mobile & Embedded

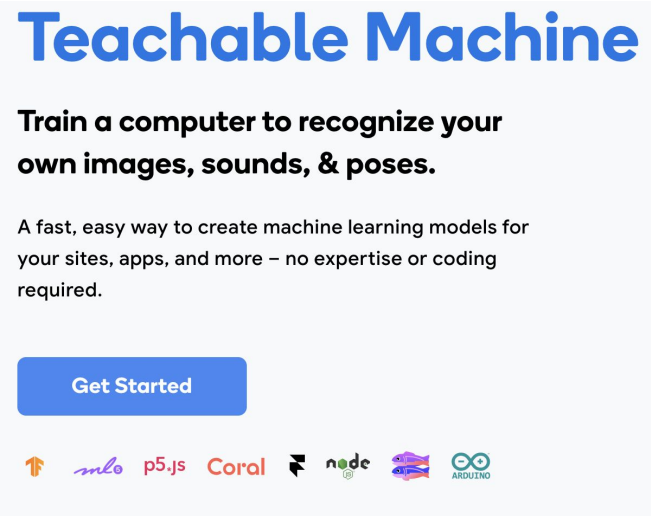



CAPTURE IMAGE
Waste type: Biodegradable
Waste Subclass: food waste, 98%
Motor 1 Status: ON
Motor 2 Status: OFF
BIODEGRADABLE

Check out the complete project here 👇

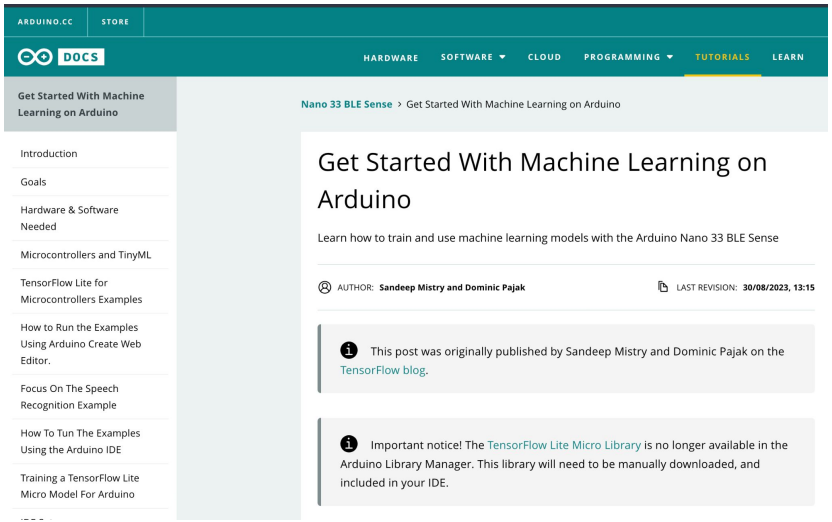
Resources for you

New to ML?



https://teachablemachine.withgoogle.com/

New to ML on edge?



https://docs.arduino.cc/tutorials/nano-33-ble-sense/get-started-with-machine-learning#introduction

Aashi Dutt

Organizer, TensorFlow User Group Chandigarh

Connect With Me 👋

@AashiDutt

Join TFUG Chandigarh Chapter 🥳

@TFUGChandigarh

@tfugchandigarh