UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
School of Professional & Executive Development

Master en Internet of Things

AI on the Edge

Barcelona, 7 de junio de 2022

About me

My name is **Sergi Mercadé Laborda,** and I am a Professional Project Engineer at Distributed Artificial Intelligence Research Area at i2CAT Foundation.

At DAI we deploy AI on the Edge PoCs and solutions for European and national research projects, public institutions and private companies.

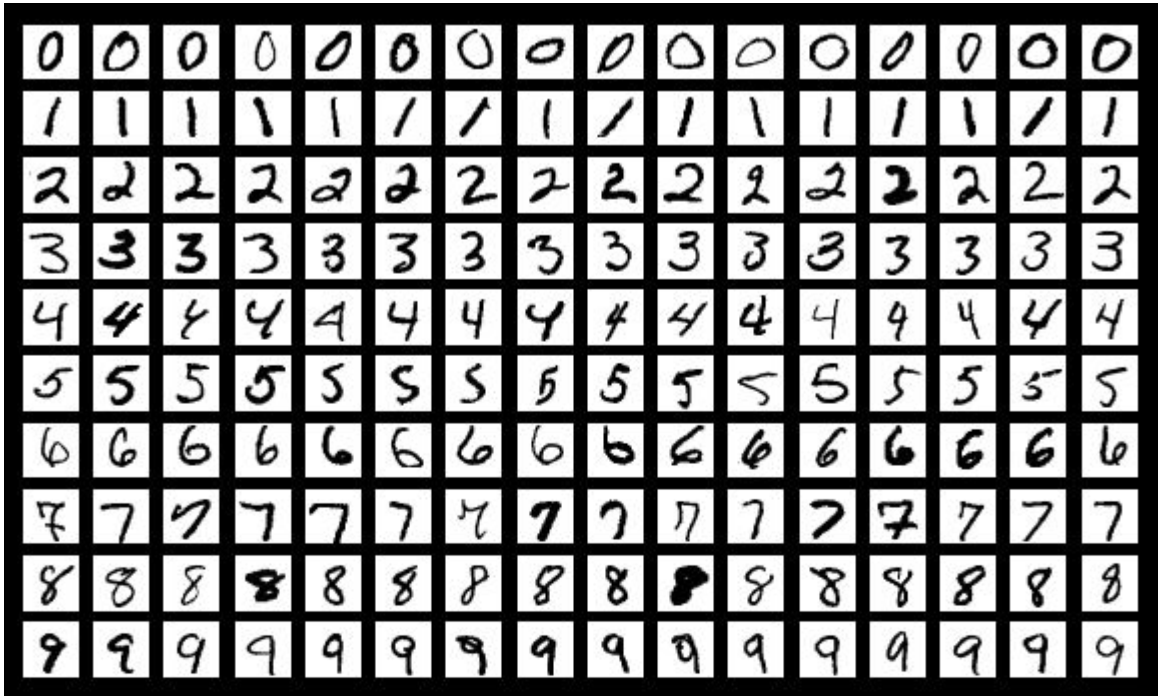
sergi.mercade@i2cat.net

Contents

- ARTIFICIAL INTELLIGENCE ON THE EDGE
  - Introduction to machine learning, Data, Neural Networks, ML Workflow, AI on the Edge and IoT applications, Edge AI devices
- GOOGLE CORAL AND TENSORFLOW
  - Introduction to Google Coral, Edge TPU, Coral requirements, Examples
- TensorFlow, Keras and TensorFlow Lite
  - Introduction to TensorFlow + Keras and TensorFlow Lite

Artificial Intelligence on the Edge

- Introduction
- Neural Networks
- ML Workflow
- AI on the Edge and IoT applications
- Edge AI devices



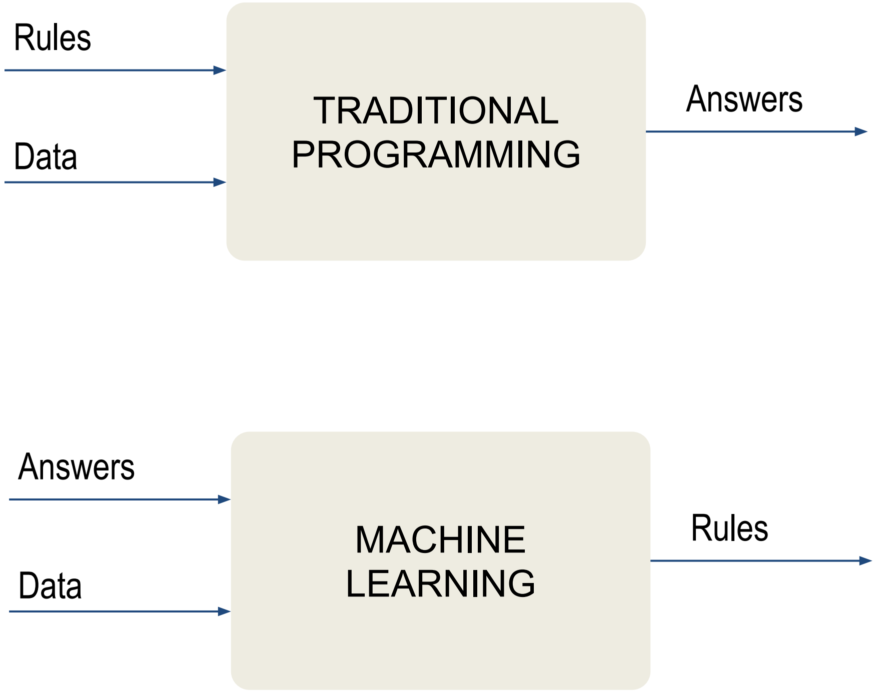
Hand-written digits recognition

This is one of the “Hello world” problems in the ML field.

Source: By Josef Steppan - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=64810040

What do we want?

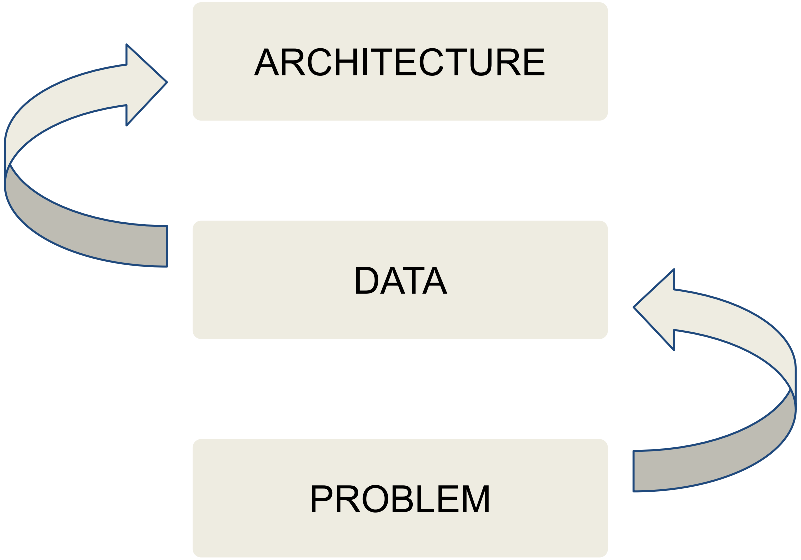
FROM DATA AND ANSWERS GET THE RULES



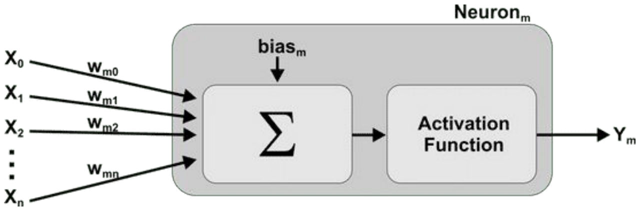
Source: Machine Learning Zero to Hero (Google I/O'19)

DATA
The fuel for AI

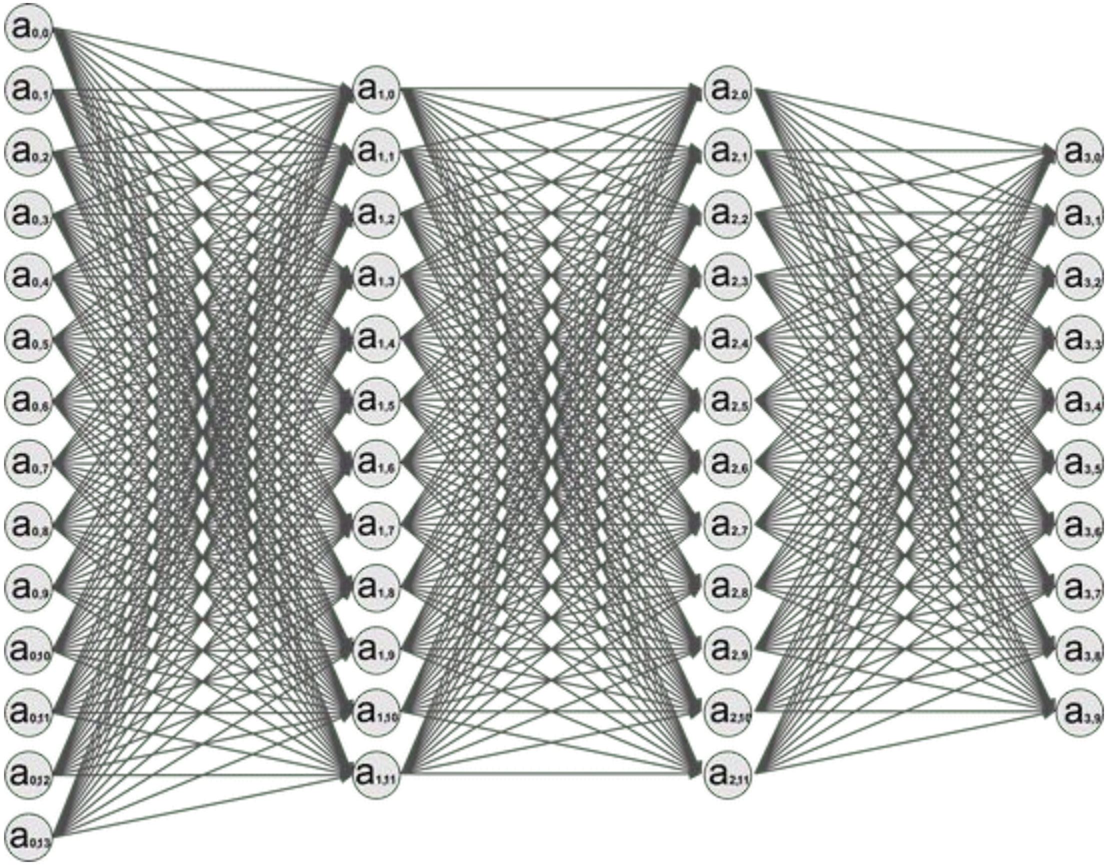
The problem you are trying to solve, your data and the architecture of your model are deeply intertwined

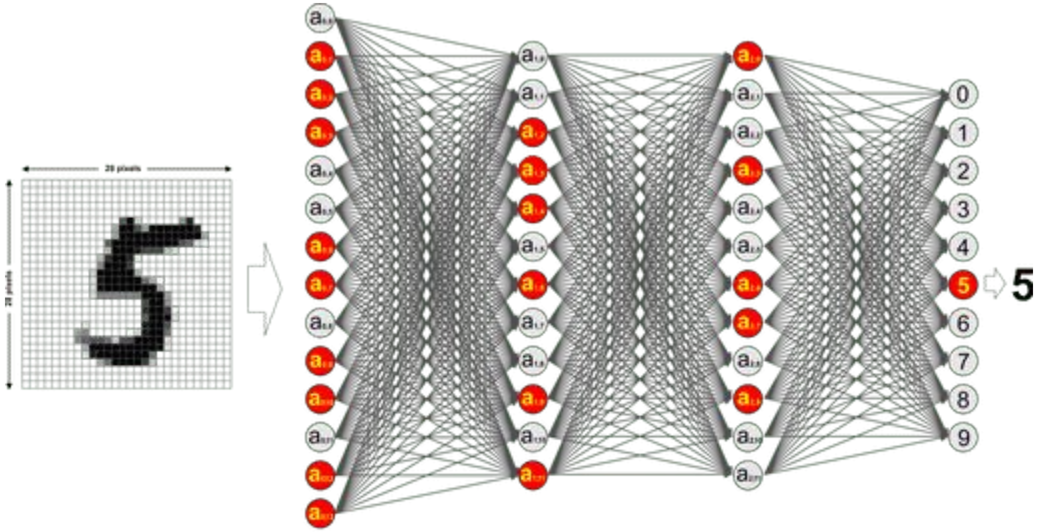


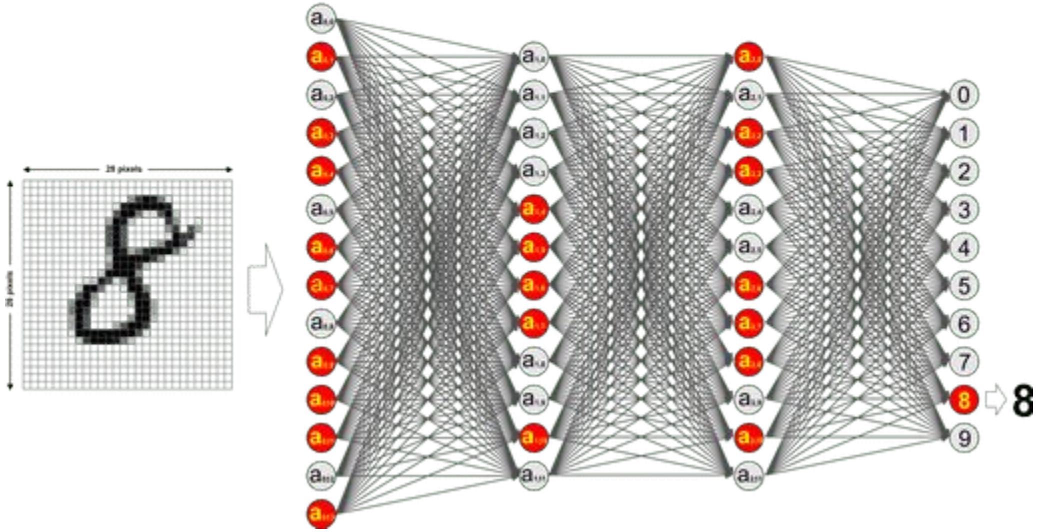
Neural Networks



Supervised learning

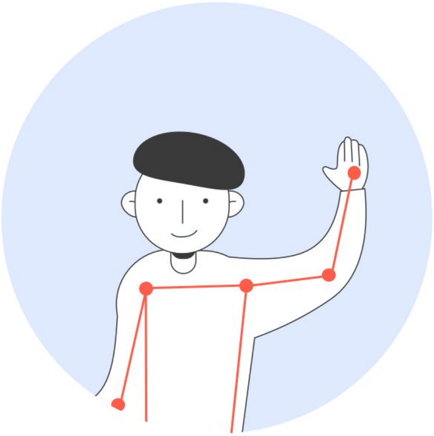


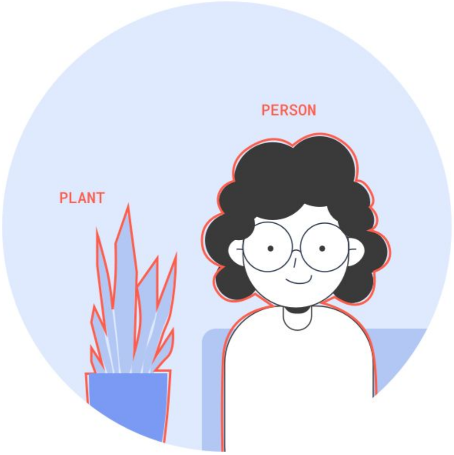

28 pixels
28 pixels
0
1
2
3
4
5
6
7
8
9
5

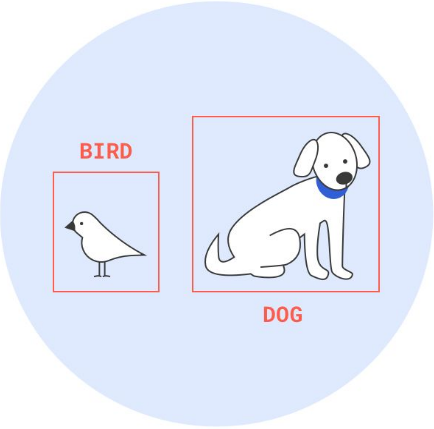

28 pixels
28 pixels
0
1
2
3
4
5
6
7
8
9
8

Some NN media-related applications

- Pose estimation
- Image segmentation
- Object detection
- Key phrase detection

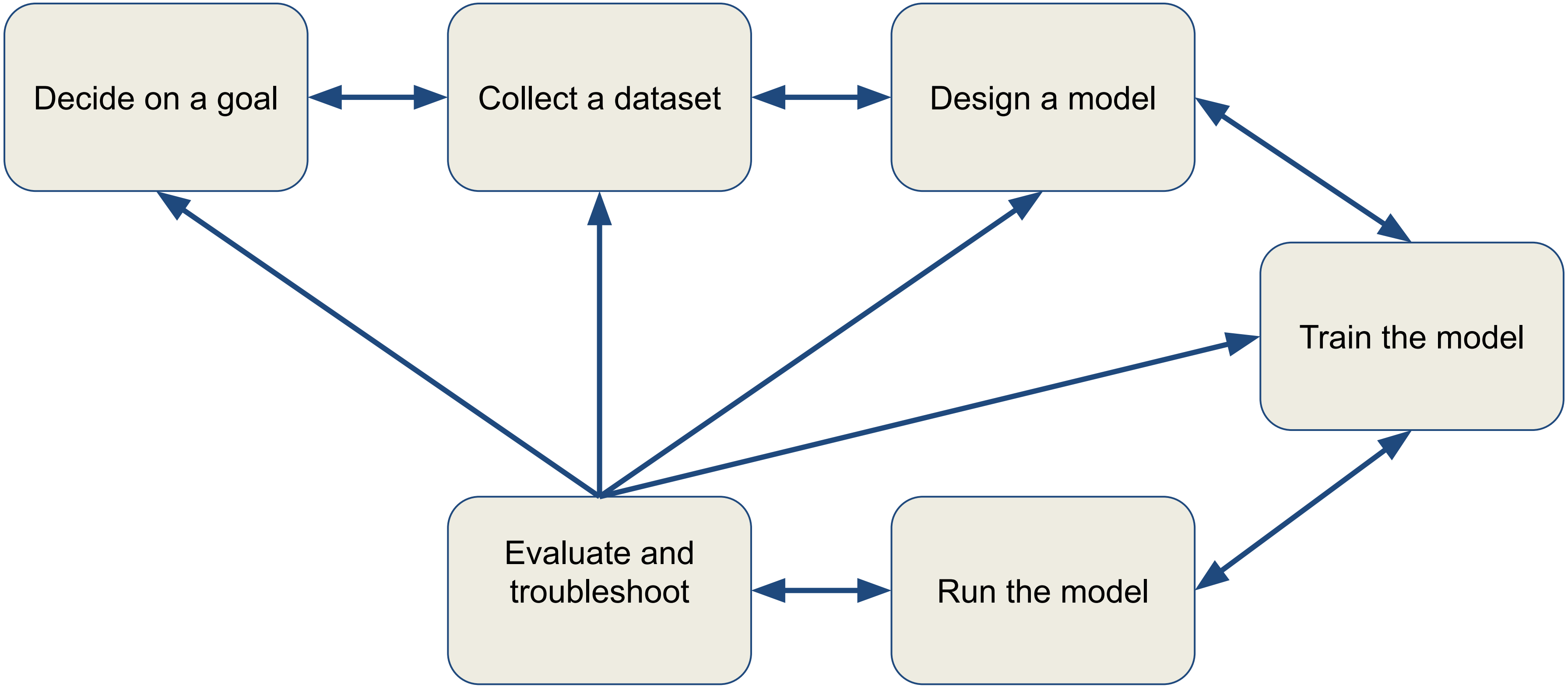






Source: coral.ai

ML Workflow



Decide on a goal

Collect a dataset

Design a model

Train the model

Evaluate and troubleshoot

Run the model

Decide on a goal

- Is there any good on using AI for the problem you are trying to solve?

- Identify the AI module inside the system pipeline. What inputs does it has? What outputs we expect to obtain?

- Hand-written digits: we want to build a classification system capable of classifying the input as a number from 0 to 9.

Collect a dataset

- We must select relevant information for the dataset.
- Deep Learning models can be robust to noise, but irrelevant data can make them vulnerable.
- The dataset must reflect the real final working environment of the system.
- This step can be one of the most expensive and/or time consuming of the workflow.

Design a model

This stage is affected by the previous ML workflow stages:

- The problem definition.
- The dataset available.
- The ways you can transform the data.
- The constraints of the target device (especially in the AI on the edge field)

It is an iterative process.

"Designing a model is both an art and a science, and model architecture is a major area of research"

Warden, Pete, Situnayake, Daniel. TinyML. O'Reilly Media.

Train the model

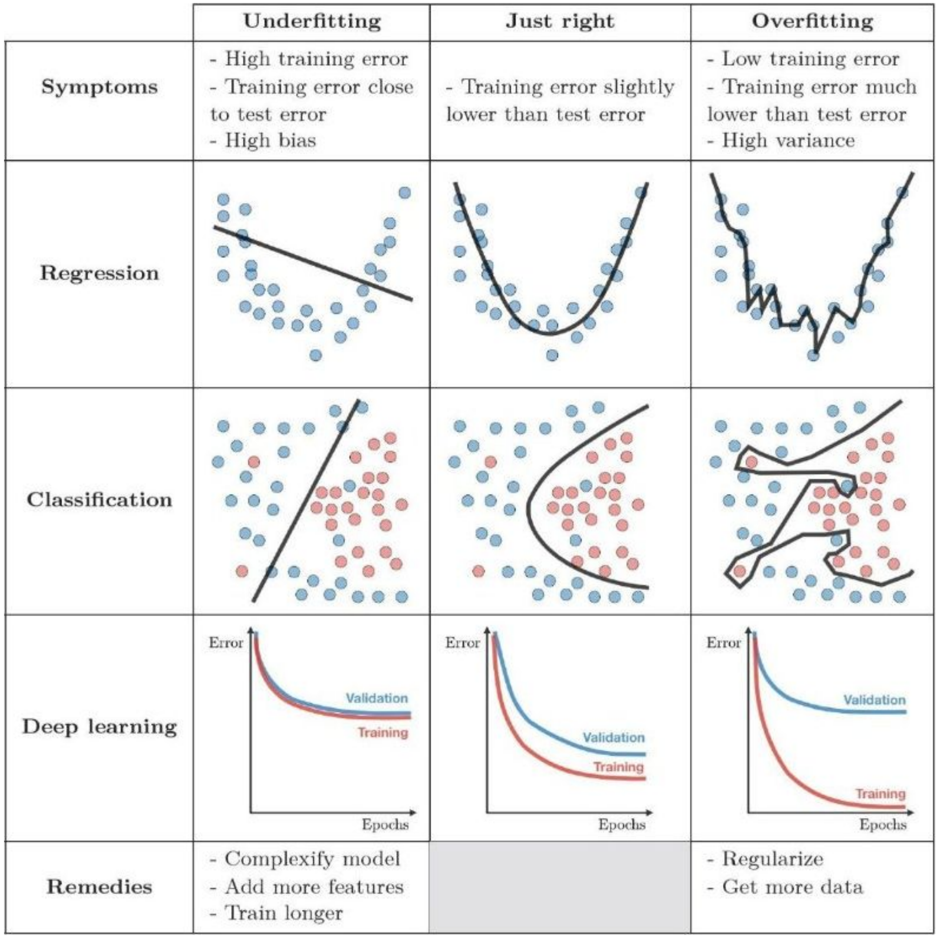
- Depending on the model's complexity, this stage can last several days or even weeks.
- The dataset must be split in 3 subsets: train, validation, test.
- Batches of data are fit into the network and the parameters are adjusted. This process is repeated for several epochs until the outputs are similar to the expected ones.
- During the training, we must monitor various metrics to make sure the model is learning as expected.

Underfitting

The model has not learned the relations in the data to make good predictions.

Overfitting

The model has learned the data too well and can't generalize over new data.



	Underfitting	Just right	Overfitting
Symptoms	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error	- Low training error - Training error much lower than test error - High variance
Regression			
Classification			
Deep learning			
Remedies	- Complexify model - Add more features - Train longer		- Regularize - Get more data

Hyperparameters

All the parameters that are not “learned” during training are hyperparameters.

The hyperparameters are adjusted considering the problem we are trying to solve and the training results.

One technique used to find a good collection of hyperparameters is called gridsearch.

Run the model

When the model achieves the desired performance, it must be implemented inside the system's pipeline and make sure the whole system works as expected.

Evaluate and troubleshoot

After deploying the model and running it on your target device, we must check that its performance is the expected one in the real-world environment.

AI on the Edge and IoT applications

AI on the Edge is the deployment of AI solutions where the data is created or on the system's periphery.

"REAL TIME AND LOCAL AI"

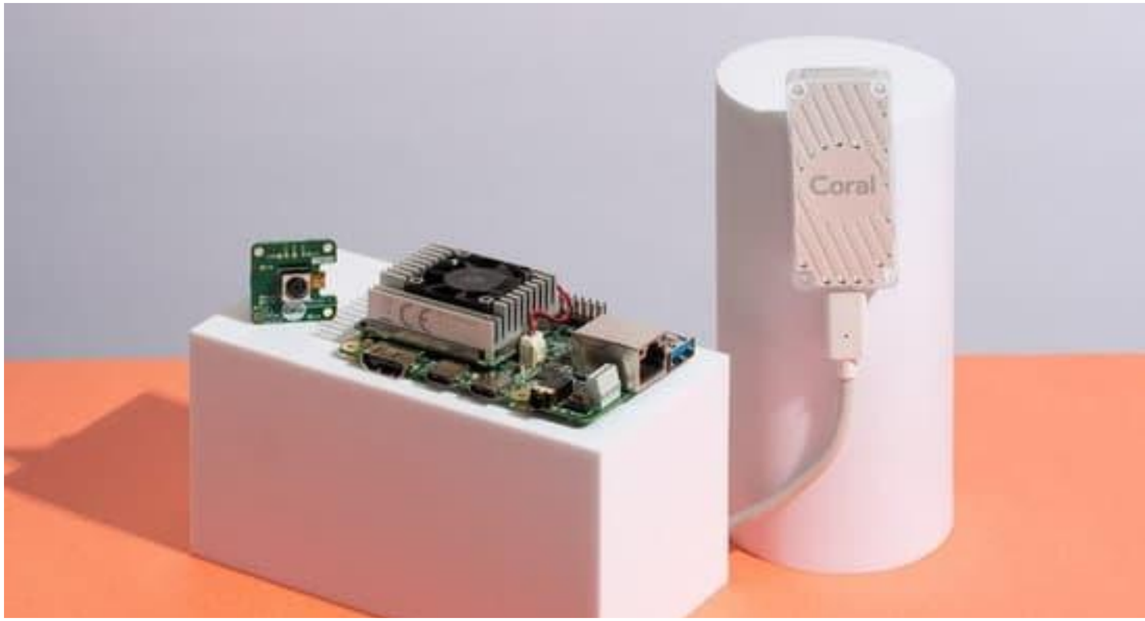
AI on the Edge examples

- An autonomous car.
- A security camera with a local people detection algorithm.
- A local classifier in a factory's robotic arm.
- A synthetic sensor that uses AI.
- The "Ok Google" keyword detection in our phones or our smart speakers

Edge AI devices

Google Coral and Tensorflow

- Introduction to Google Coral
- Coral Requirements
- Examples
- TensorFlow and Keras
- TensorFlow Lite



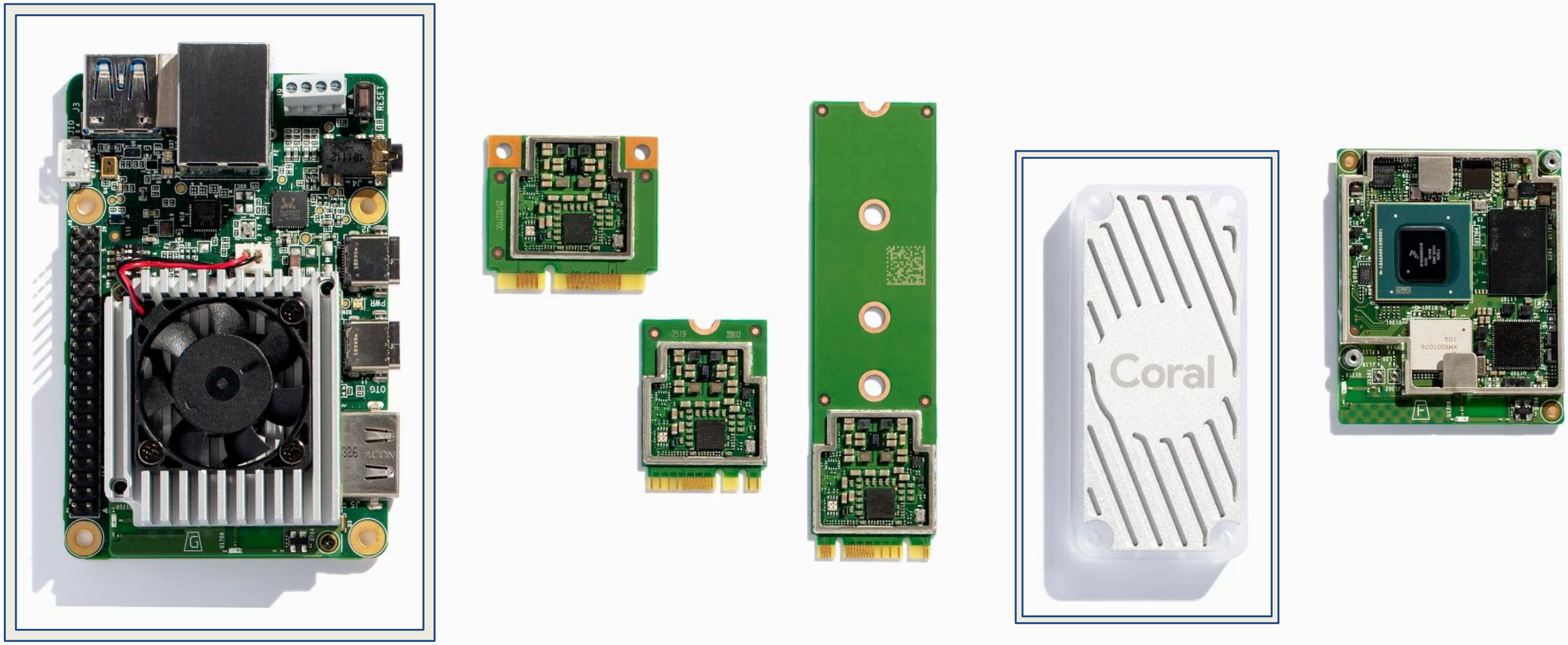
Introduction to Google Coral

Coral is a hardware and software platform for building intelligent devices with fast neural network inferencing.

Efficient **Fast**

Offline **Private**

Coral’s family

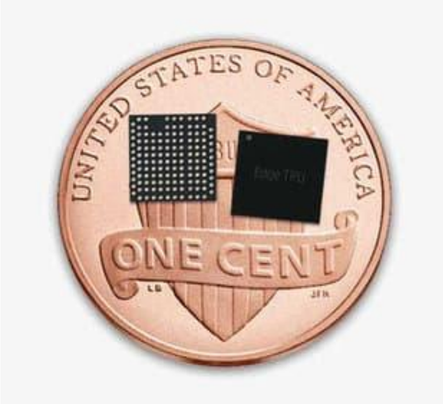


Coral Dev Board

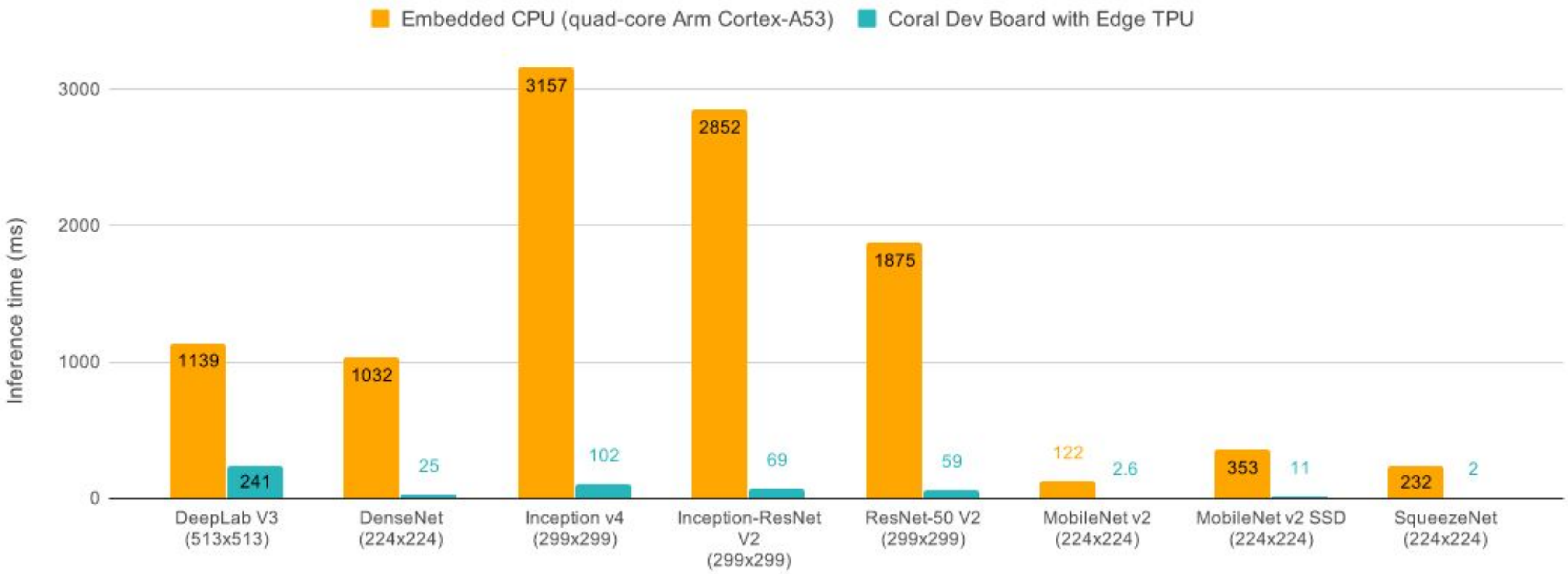
Coral USB accelerator

Edge TPU

Small ASIC built by Google that's specially-designed to execute state-of-the-art neural networks at high speed, with a low power cost.



It can perform 4 tera-operations per second (TOPS) using 0.5 watts per TOP.

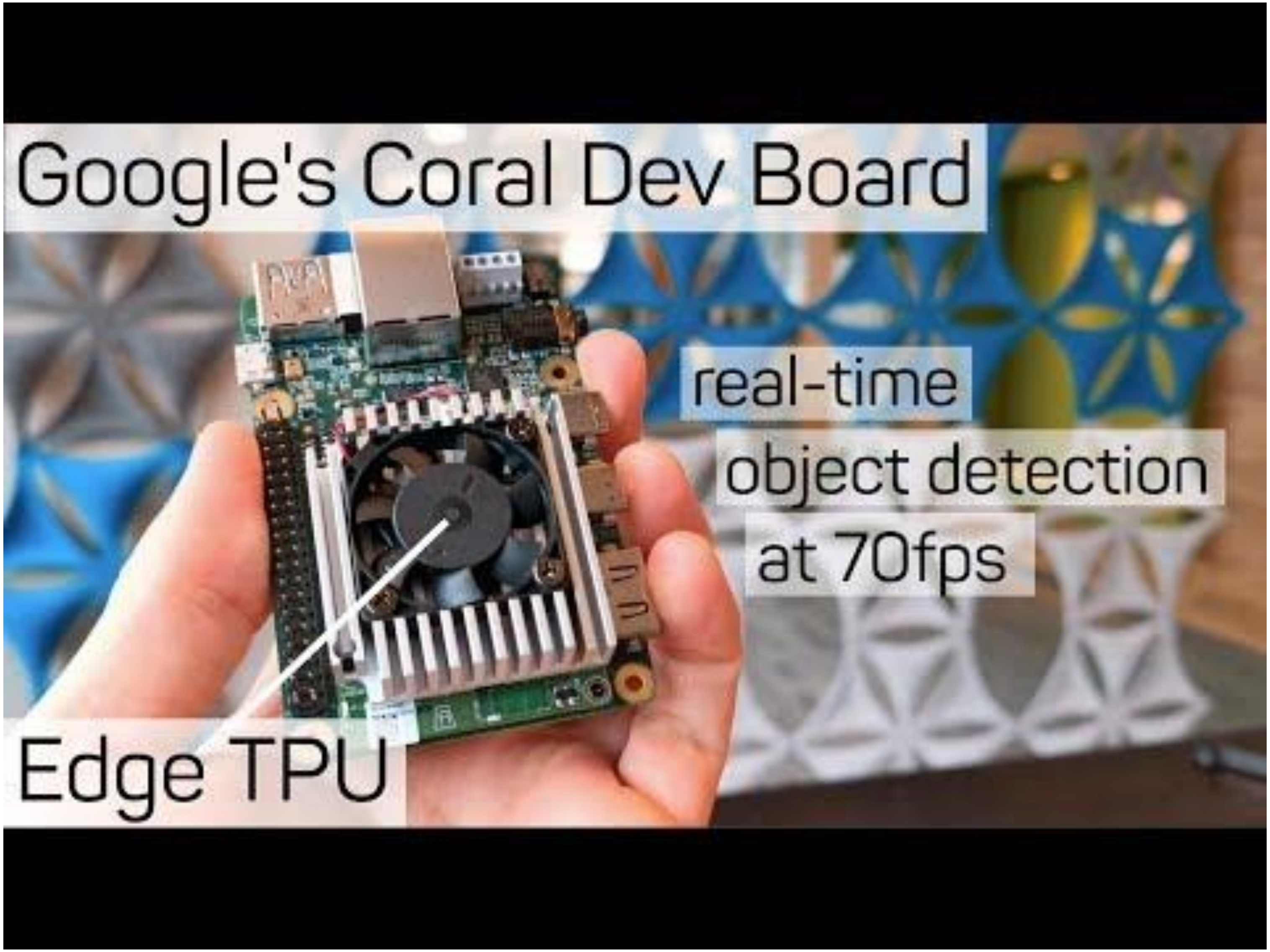


Source: https://coral.ai/technology

Edge TPU performance

Model architecture	Desktop CPU*	Desktop CPU* + USB Accelerator (USB 3.0) with Edge TPU	Embedded CPU**	Dev Board*** + with Edge TPU
MobileNet v1	47 ms	2.2 ms	179 ms	2.2 ms
MobileNet v2	45 ms	2.3 ms	150 ms	2.5 ms
Inception v1	92 ms	3.6 ms	406 ms	3.9 ms
Inception v4	792 ms	100 ms	3,463 ms	100 ms

*Desktop CPU: 64 bit Intel(R) Xeon(R) E5 1650 v4 @ 3.60GHz
**Embedded CPU: Quad-core Cortex-A 53 @ 1.5 GHz
***Dev Board: Quad-core Cortex-A53 @ 1.5 GHz + Edge TPU
Source: Introducing Google Coral: Building On-Device AI (Google I/O'19)


Google's Coral Dev Board
real-time
object detection
at 70fps
Edge TPU

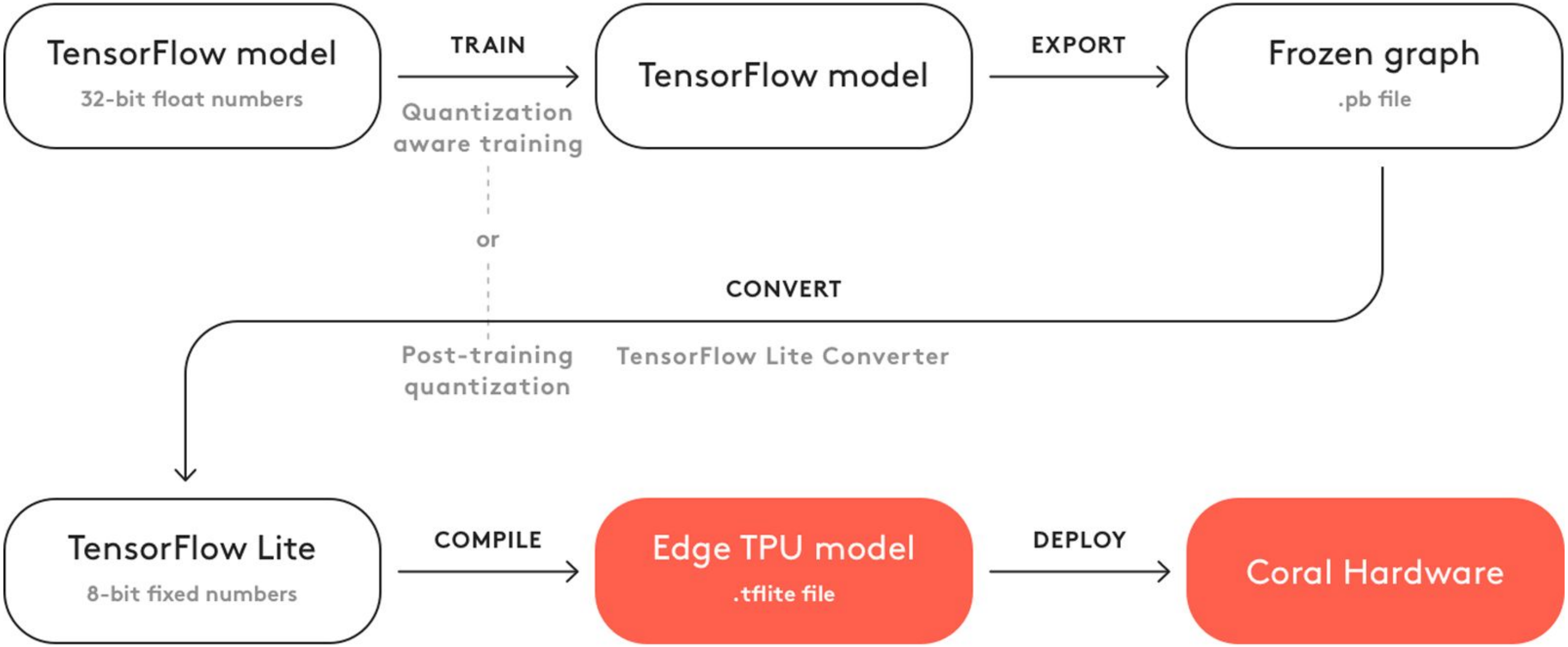
Coral Requirements

To provide high-speed neural network performance with a low-power cost, the Edge TPU supports a specific set of neural network operations and architectures.

It supports only TensorFlow Lite models that are fully 8-bit quantized and then compiled specifically for the Edge TPU.

Source: https://coral.ai/docs/edgetpu/models-intro#compatibility-overview

Coral Requirements



Source: https://coral.ai/docs/edgetpu/models-intro#compatibility-overview

- Tensor parameters are quantized (8-bit fixed-point numbers; int8 or uint8).
- Tensor sizes are constant at compile-time (no dynamic sizes).
- Model parameters (such as bias tensors) are constant at compile-time.
- Tensors are either 1-, 2-, or 3-dimensional. If a tensor has more than 3 dimensions, then only the 3 innermost dimensions may have a size greater than 1.
- The model uses only the operations supported by the Edge TPU.

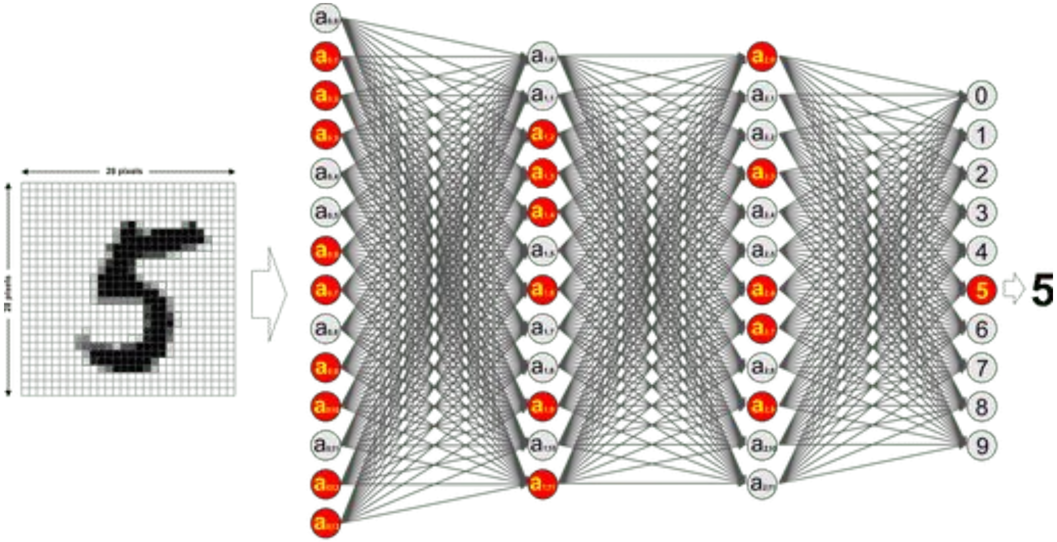
Source: https://coral.ai/docs/edgetpu/models-intro#compatibility-overview

Example

https://coral.ai/projects/teachable-sorter/

https://teachablemachine.withgoogle.com/

TensorFlow and Keras introduction



TensorFlow and Keras

TensorFlow is an open source library designed by Google to train and develop ML models. It is used by researchers and developers alike.

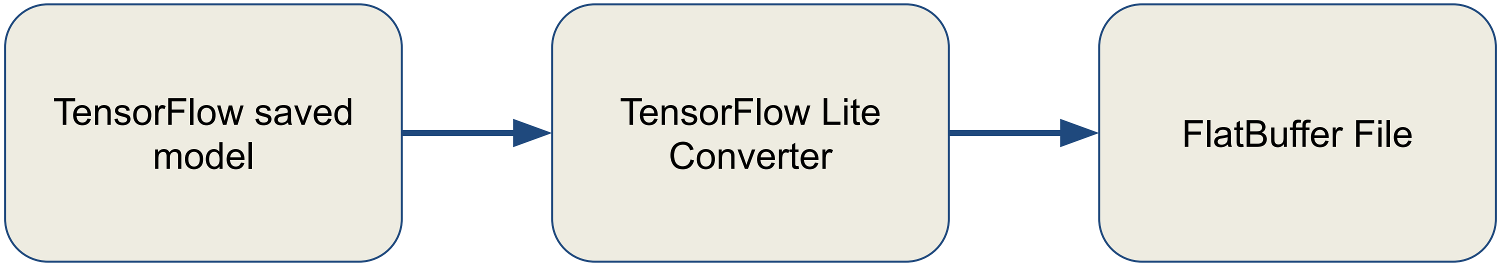
Keras is a high-level API that greatly simplifies the development of ML models.

TensorFlow Lite

TensorFlow Lite is a set of tools for running TensorFlow models on “edge devices”

- TensorFlow Lite Converter: Converts TensorFlow models into special efficient models for constrained devices.

- TensorFlow Lite Interpreter: Efficiently runs TensorFlow Lite models.



TensorFlow saved model

TensorFlow Lite Converter

FlatBuffer File

Resources

- 3blue1brown YouTube videos (Introduction to NN math)
- Sebastian Raschka, Vahid Mirjalili. Python Machine Learning. Marcombo. (Introduction to ML math and TF+Keras)
- https://www.tensorflow.org
- https://coral.ai
- Warden, Pete,Situnayake, Daniel. TinyML. O'Reilly Media. (TF and TF Lite for Edge devices)