ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

APAI Lab3: DNN Quantization

Lorenzo Lamberti, Francesco Conti, Alberto Dequino

Luka Macan, Nazareno Bruschi, Alessio Burrello, Davide Nadalini. *(University of Bologna)*

lorenzo.lamberti@unibo.it

f.conti@unibo.it.

In this Hands-on session:

A first-time user of Pytorch framework will learn how to :

- shrink a NN, by acting on the number of layers, channels, or stride factor
- Quantize a NN down to 2 bits
- Use netron to visualize a ONNX representation of a CNN

Tasks:

1. Load model's trained weights of LAB1;
2. Reduce network's size under 5MMAC;
3. Re-train the reduced network and verify network's accuracy;
4. Quantize with QuantLab;
5. Export Onnx and analyze the float32 and quantized models with Netron.

All the details about the tasks are explained in the pdf document attached.

Colab: https://github.com/EEESlab/APAI-LAB03-DNN-Shrinking-and-Quantization

How to deliver the assignment

- Use Virtuale platform to load your file
- update only the .ipynb file, **named as follows**: LAB<lab_number>_APAI_<yourname>.ipynb

Important: the notebook must be pre-run by you. Outputs must be correct and visible when you download it.

Assignment 1 (due 25/11/2021)

Submission status

Submission status	Draft (not submitted)
Grading status	Not graded
Due date	Thursday, 25 November 2021, 4:00 PM
Time remaining	9 days 5 hours
Last modified	Tuesday, 16 November 2021, 10:56 AM
File submissions	name of the file LAB1_APAI_Lorenzo_Lamberti.ipynb 16 November 2021, 10:56 AM
Submission comments	Comments (0)

edit submitted files

Edit submission | Remove submission

You can still make changes to your submission.

submit

Submit assignment

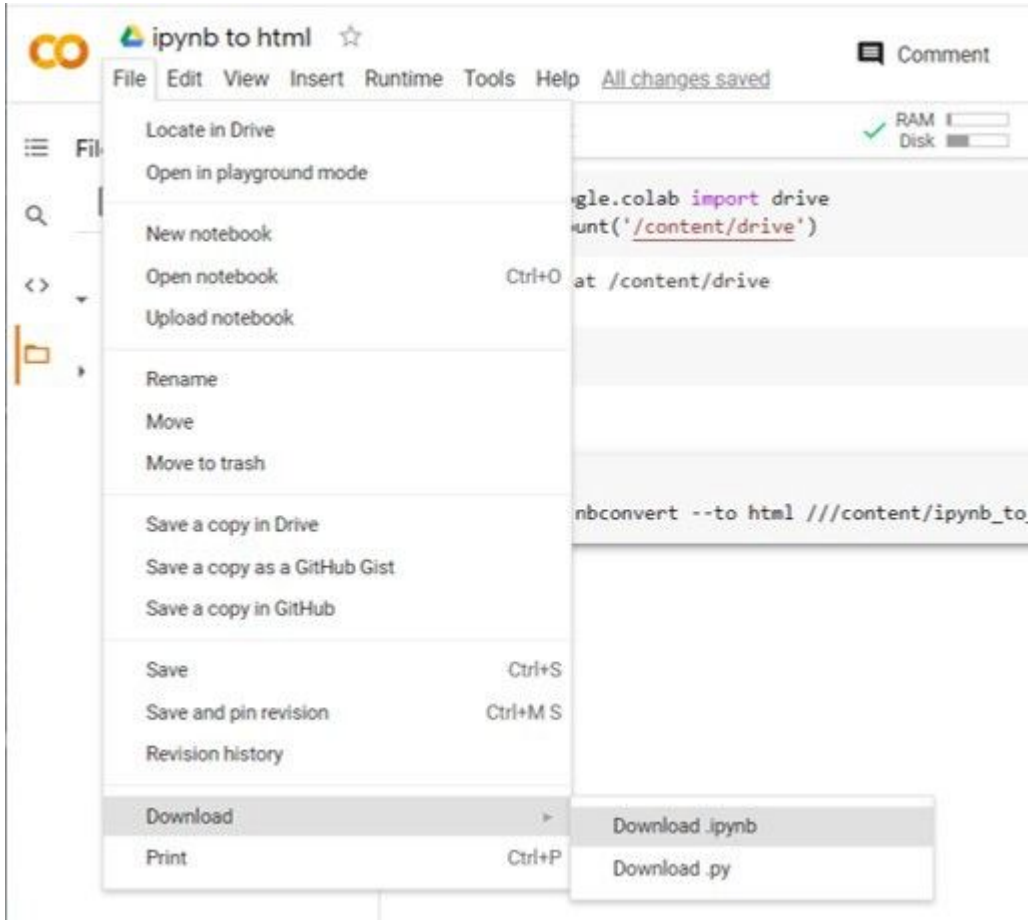
LAB DEADLINE:

10/11/2022 at 23:59

(2 weeks from today)

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

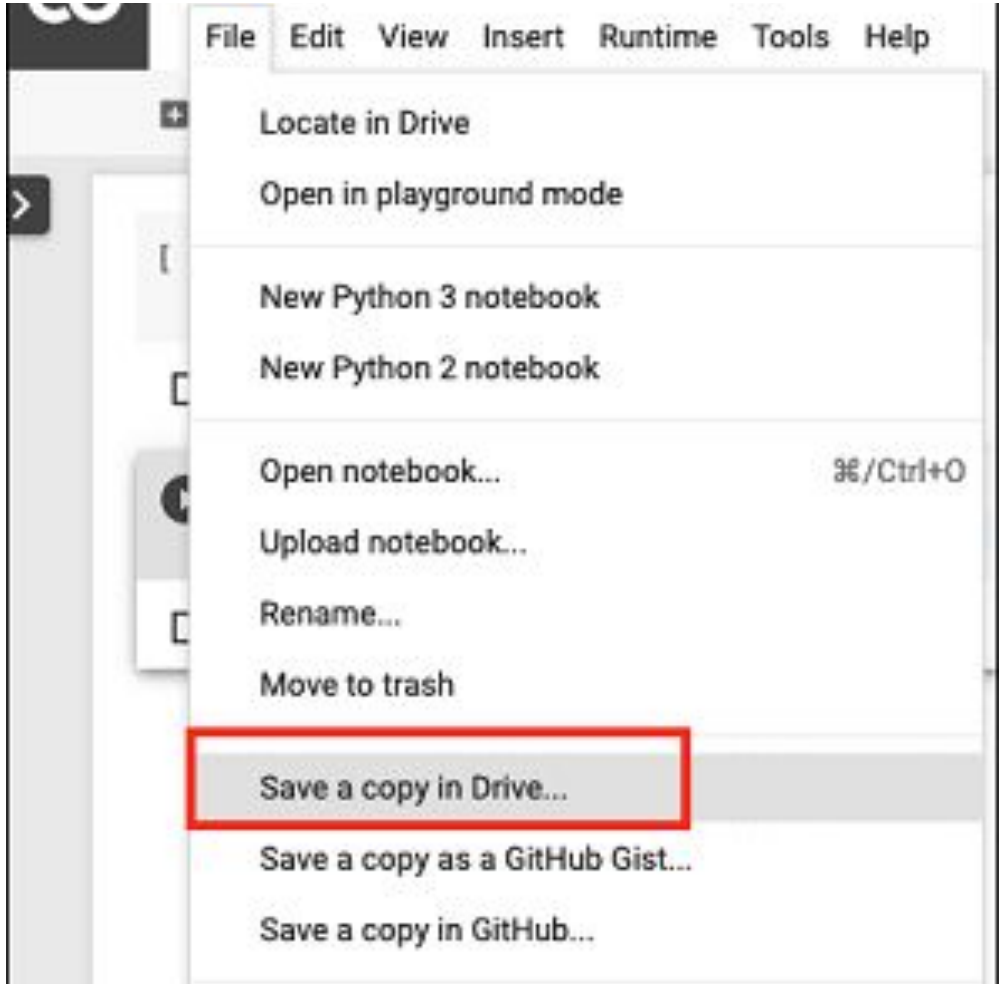
How to download the .ipynb file



Setup

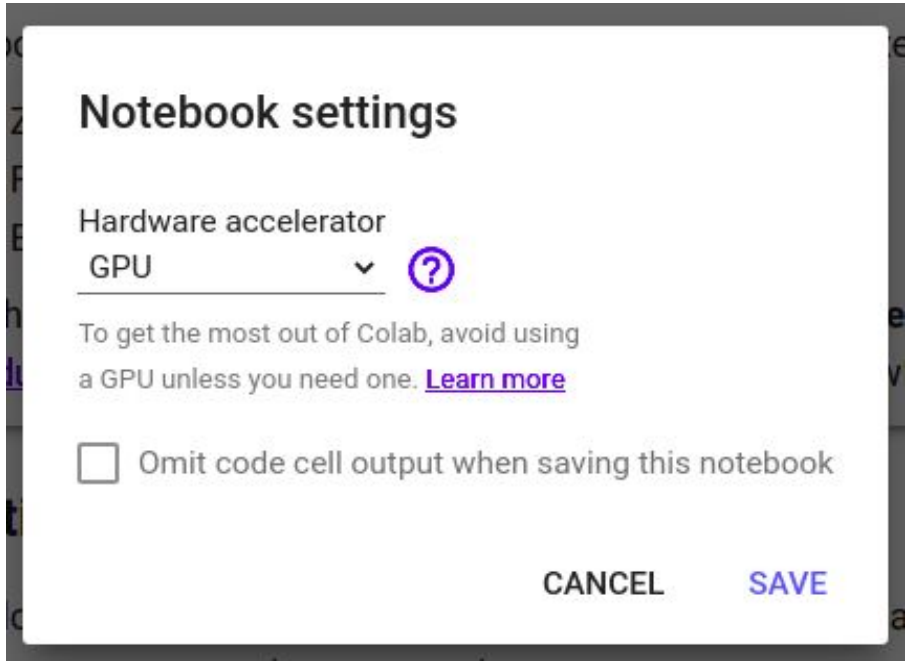
IMPORTANT:

Create your own copy of the COLAB notebook!



Others:

- Activate/deactivate GPU: Runtime -> Change runtime type
- **Note:** If you use for too much time the GPU, your account will be limited to CPU for 24h.



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

The LAB starts now !