ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

APAI Lab6: DNN Quantization

Lorenzo Lamberti, Francesco Conti, Davide Nadalini, Alessio Burrello, Nazareno Bruschi, Alberto Dequino.
(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:

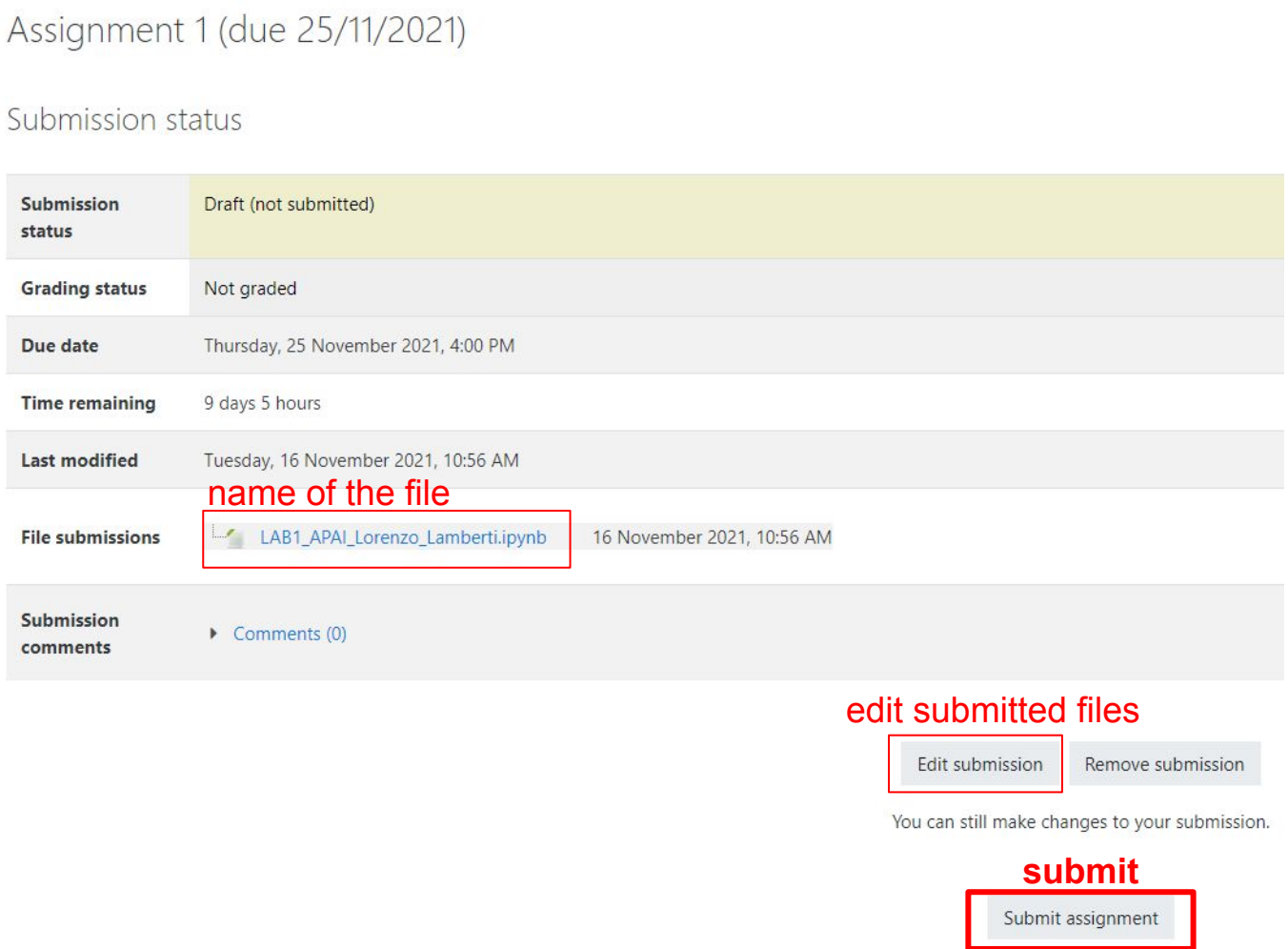
1. Load model's trained weights of LAB1;
2. Reduce network's size under 5MMAC;
3. Re-train the reduced network and verify network's accuracy;
4. Quantize with QuantLab;
5. Export Onnx and analyze the float32 and quantized models with Netron.

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

How to deliver the assignment

- Use Virtuale platform to load your file: link
- update only the .ipynb file, **named as follows**: LAB6_APAI_yourname.ipynb

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

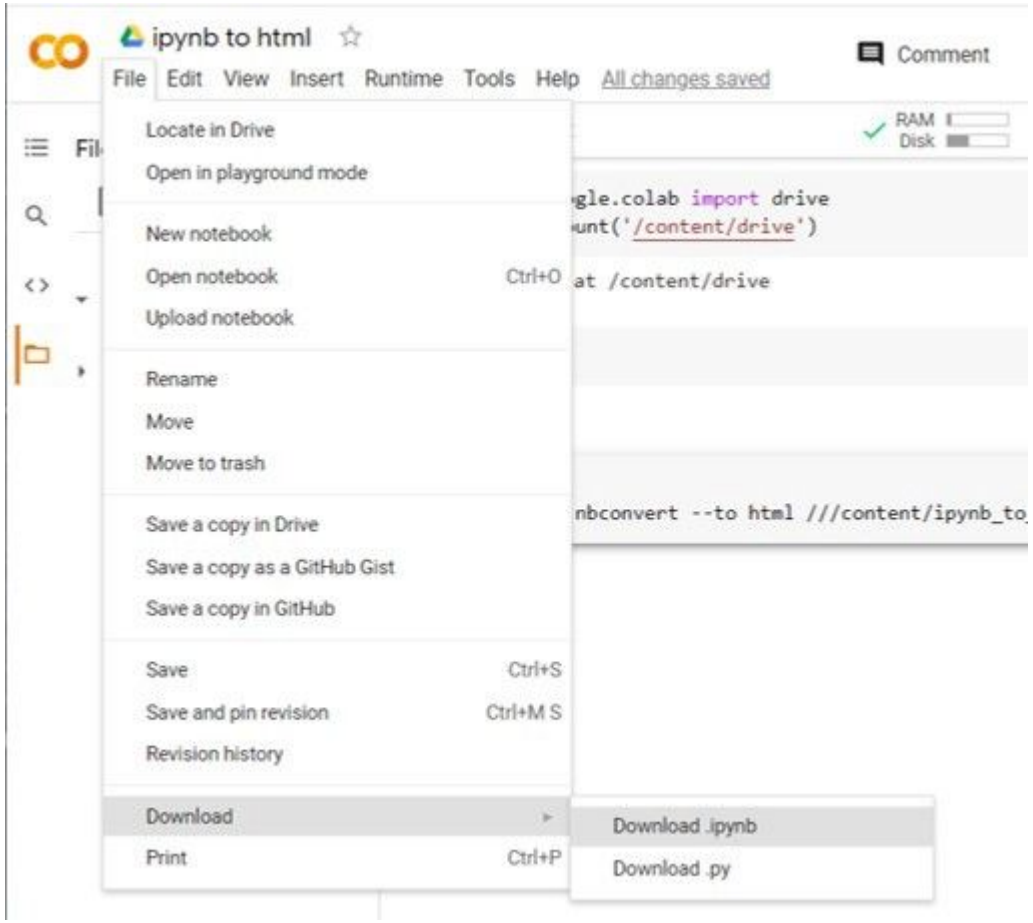


LAB6 DEADLINE:

02/12/2021 at 4PM

(1 week from today)

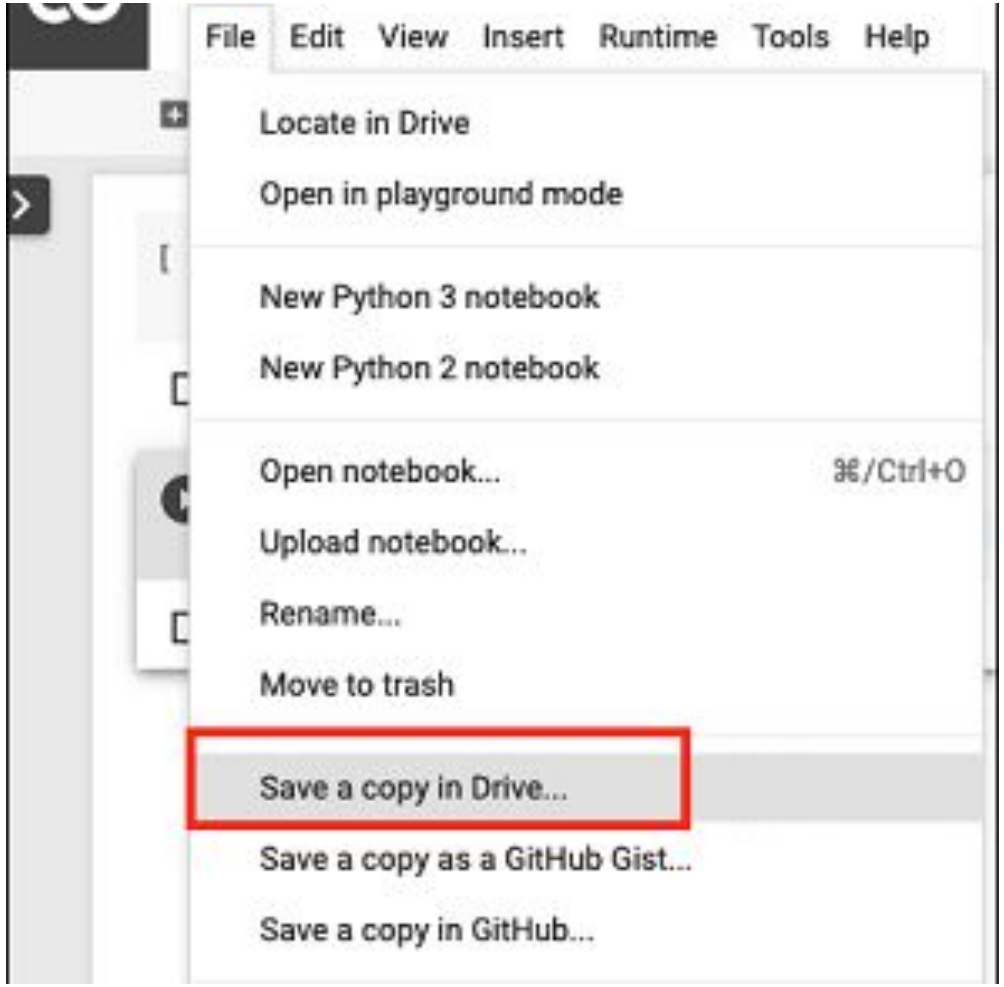
How to download the .ipynb file



Setup

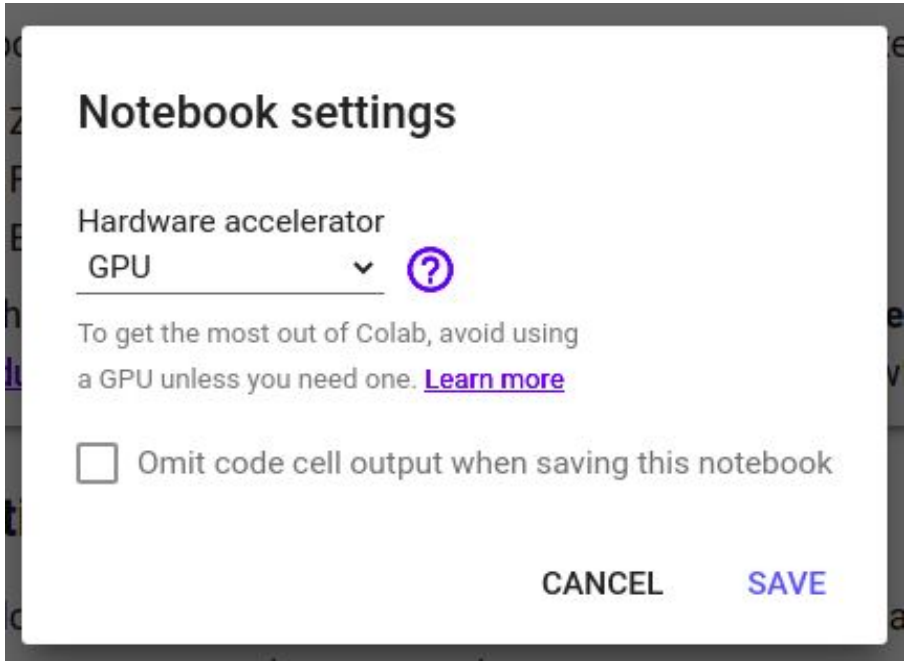
IMPORTANT:

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 !