ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

APAI Lab2: DNN Quantization

Lorenzo Lamberti, Davide Nadalini, Luca Bompani, Luka Macan, Alberto Dequino, Francesco Conti.

(University of Bologna)

lorenzo.lamberti@unibo.it d.nadalini@unibo.it luka.macan@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:

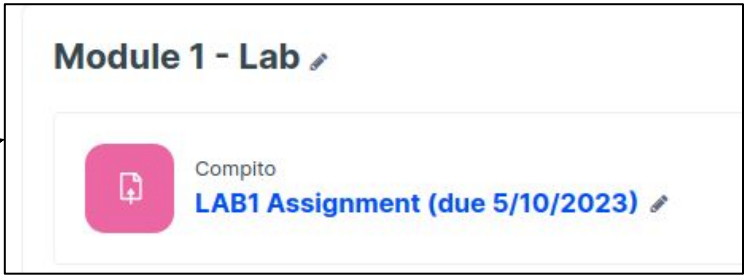
1. Load model’s trained weights of LAB1;
2. Reduce network’s size under 5MMAC;
3. Re-train the reduced network and verify network’s accuracy;
4. Quantize with QuantLab;
5. Export Onnx and analyze the float32 and quantized models with Netron.

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
- update only the .ipynb file, **named as follows**:

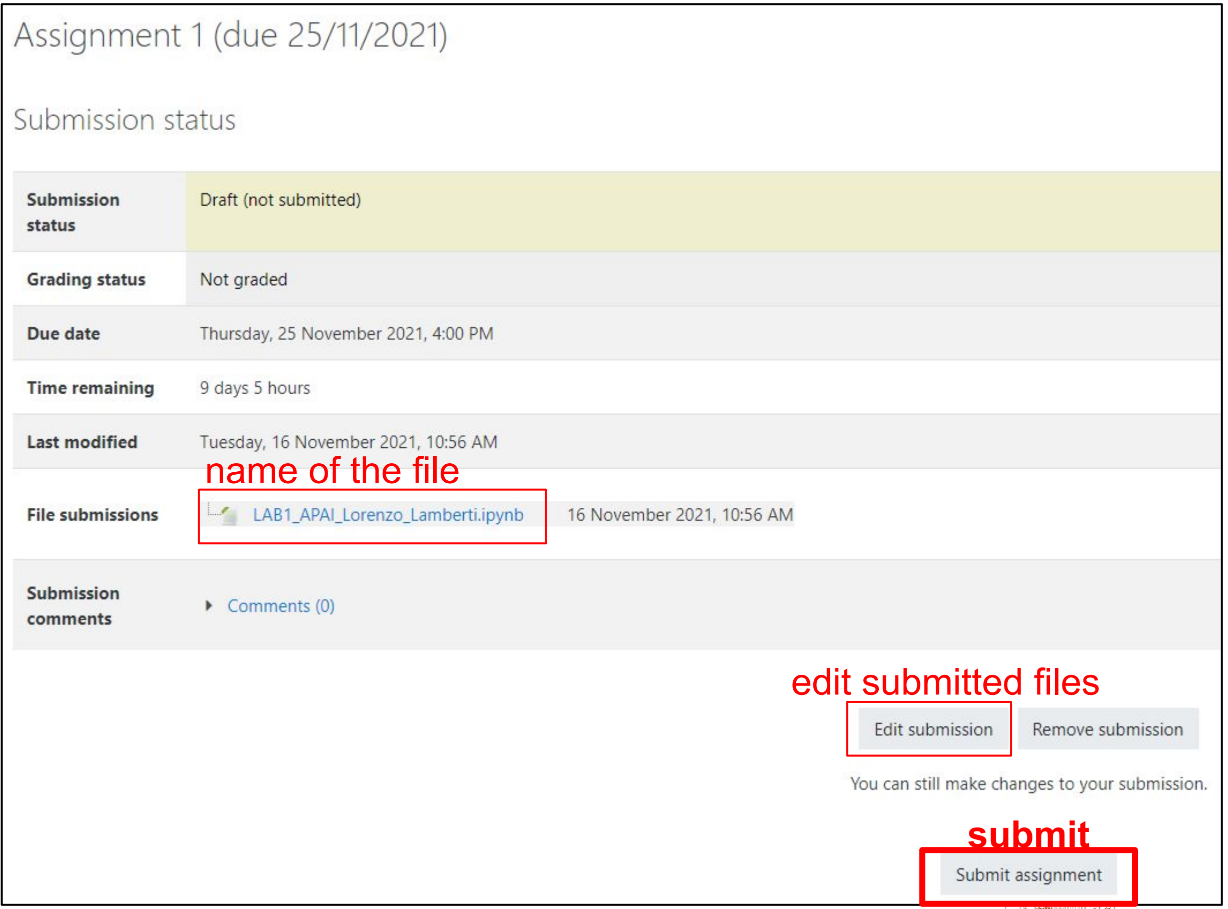
  LAB1_APAI_yourname.ipynb



Module 1 - Lab

Compito

LAB1 Assignment (due 5/10/2023)



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	LAB1_APAI_Lorenzo_Lamberti.ipynb 16 November 2021, 10:56 AM
Submission comments	▸ Comments (0)

name of the file

edit submitted files

You can still make changes to your submission.

submit

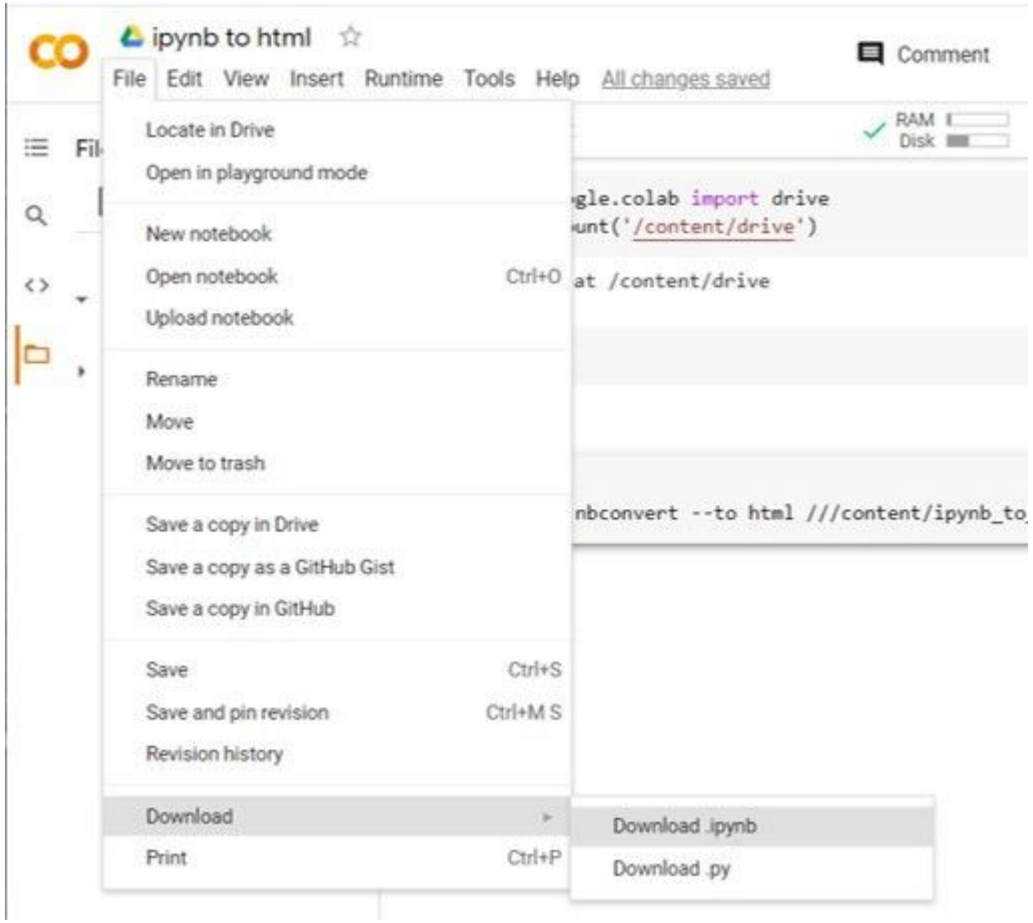
Important: the notebook must be pre-run by you. Outputs must be correct and visible when you download it.

LAB1 DEADLINE:

19/10/2023 at 16:00

(1 week from today)

How to download the .ipynb file



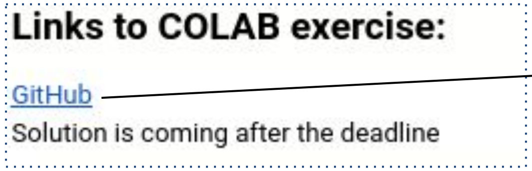
```
gle.colab import drive
unt('/content/drive')

at /content/drive

nbconvert --to html ///content/ipynb_to
```

Setup

1



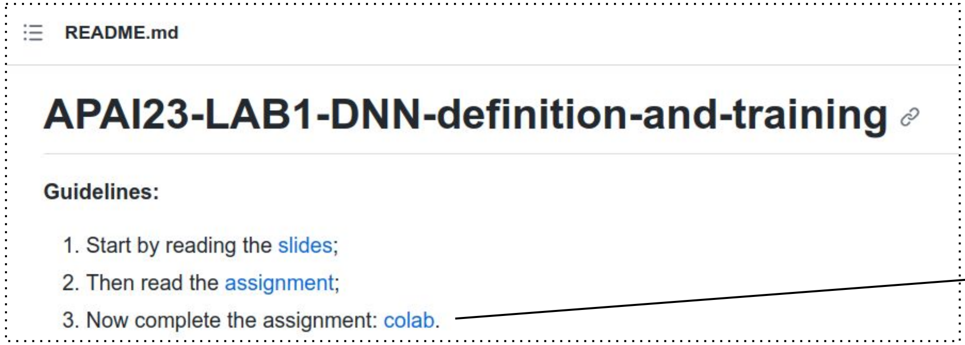
Links to COLAB exercise:

GitHub

Solution is coming after the deadline

Open github link

2



README.md

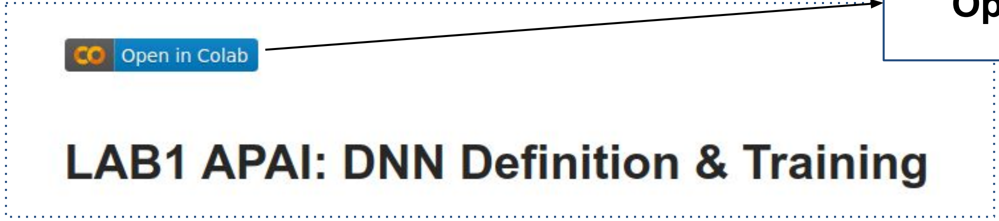
APAI23-LAB1-DNN-definition-and-training

Guidelines:

1. Start by reading the slides;
2. Then read the assignment;
3. Now complete the assignment: colab.

Open Jupyter notebook

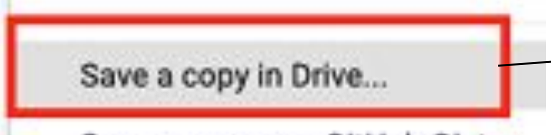
3



LAB1 APAI: DNN Definition & Training

Open In COLAB to modify it !

4

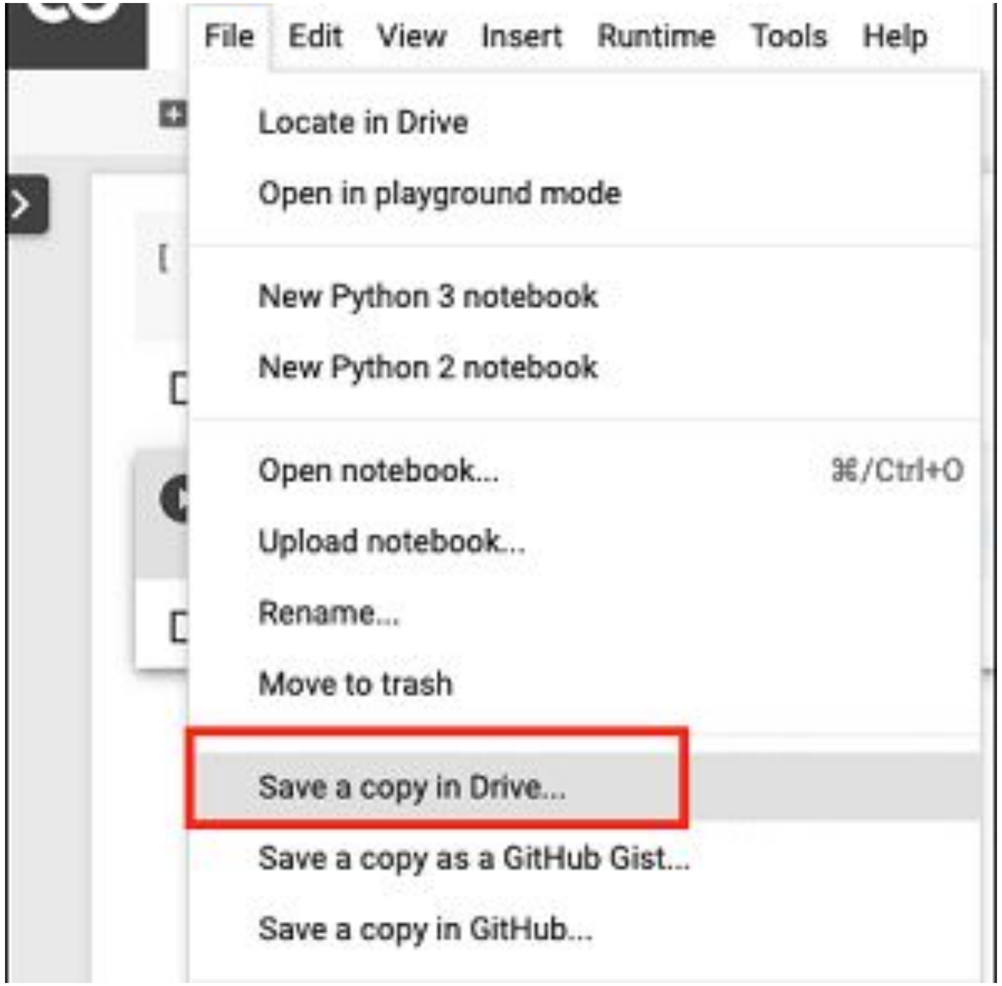


create your own copy in COLAB to modify it !

more details on step [4] of the setup

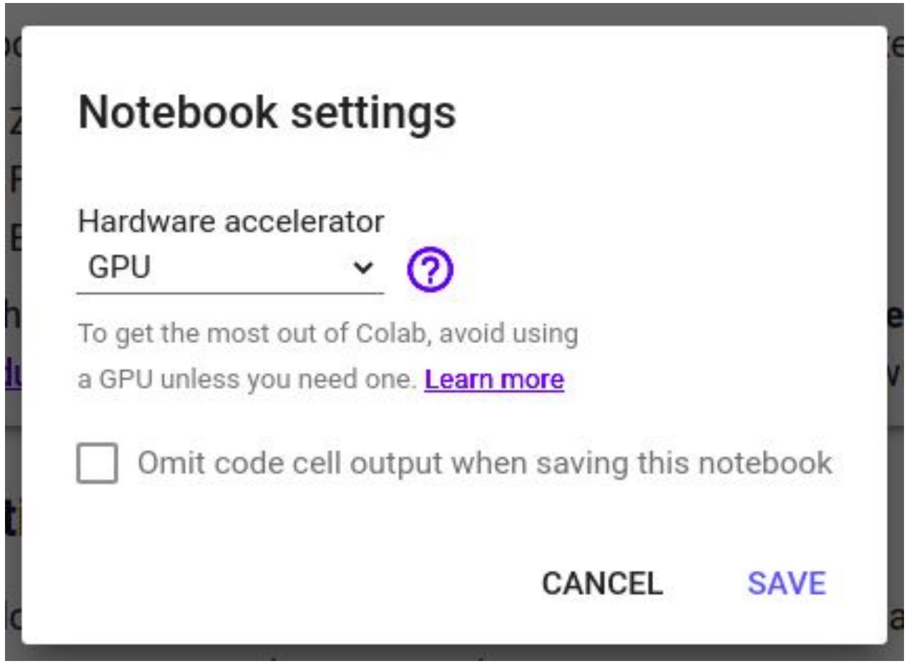
IMPORTANT:

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.



Notebook settings

Hardware accelerator
GPU

To get the most out of Colab, avoid using a GPU unless you need one. Learn more

□ Omit code cell output when saving this notebook

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

The LAB starts now !