ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

APAI Lab03: DNN Shrinking and Quantization

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

(University of Bologna)

d.nadalini@unibo.it lorenzo.lamberti@unibo.it
alberto.dequino@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:

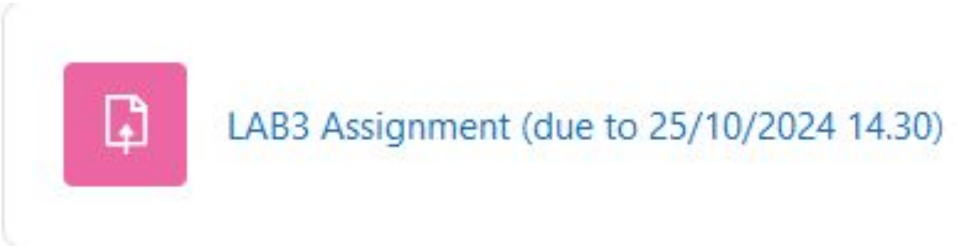
1. Load model's trained weights of LAB1;
2. Reduce network's size under 5 MMAC;
3. Re-train the reduced network and verify network's accuracy;
4. Quantize with QuantLab;
5. Export Onnx and analyze the float32 and quantized models with Netron.

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

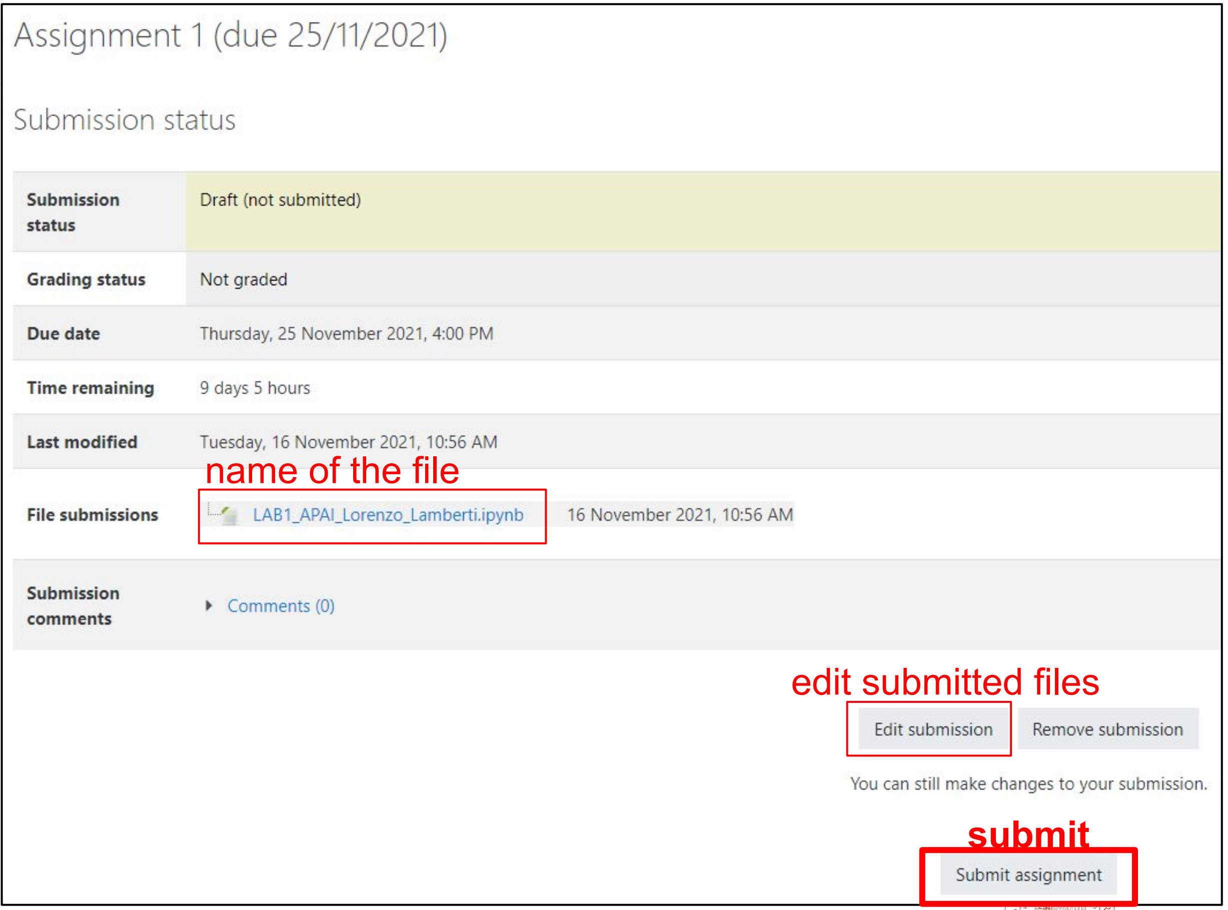
How to deliver the assignment

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

  LAB1_APAI_yourname.ipynb



LAB3 Assignment (due to 25/10/2024 14.30)



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)

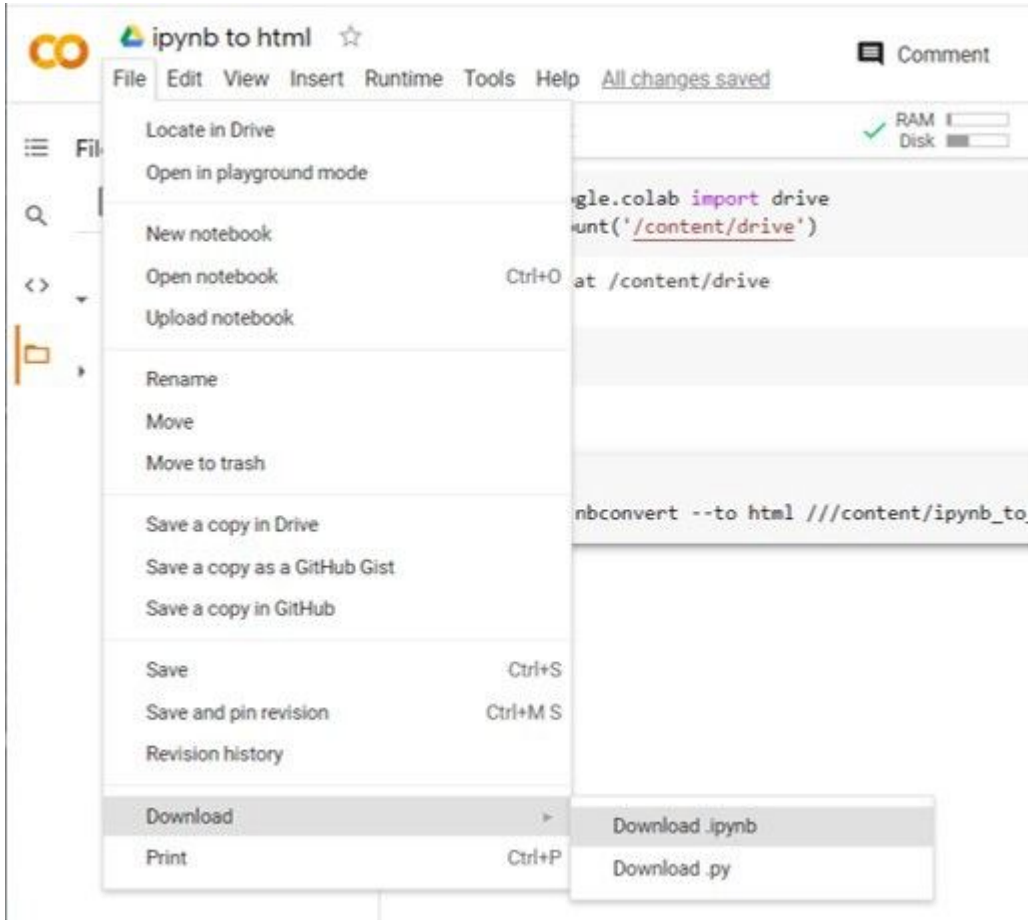
You can still make changes to your submission.

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

LAB1 DEADLINE:

25/10/2024 at 14:30

How to download the .ipynb file



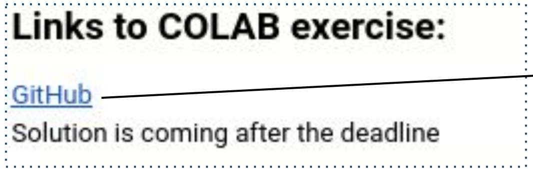
```
gle.colab import drive
unt('/content/drive')
```

```
at /content/drive
```

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

Setup

1



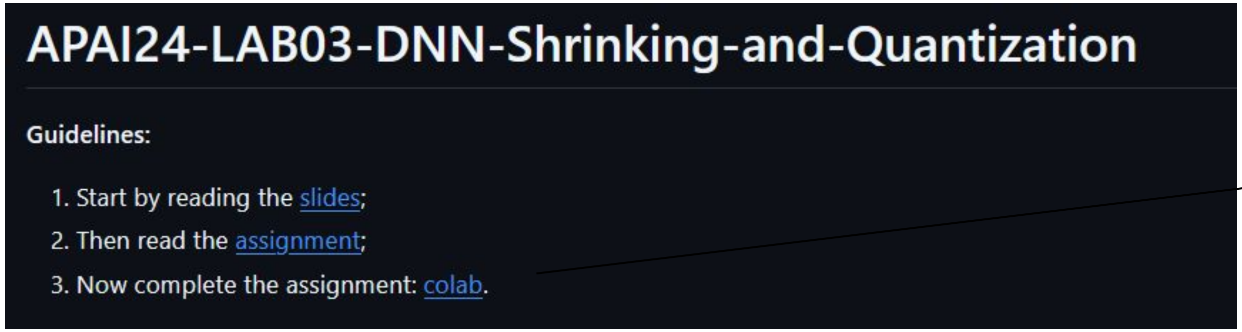
Links to COLAB exercise:

GitHub

Solution is coming after the deadline

Open github link

2



APAI24-LAB03-DNN-Shrinking-and-Quantization

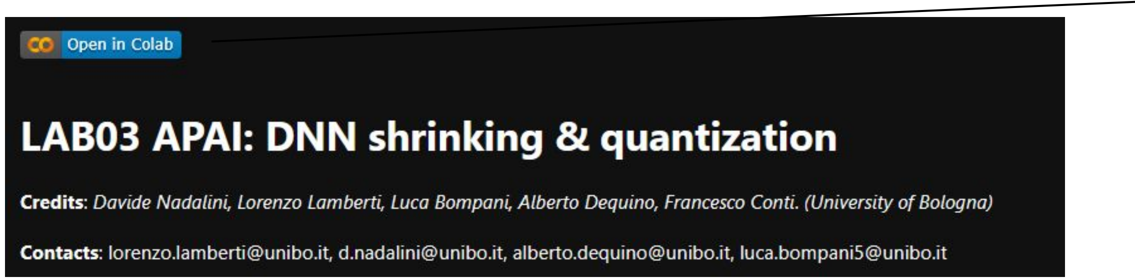
Guidelines:

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

Open Colab Jupyter notebook

Open In COLAB to modify it !

3

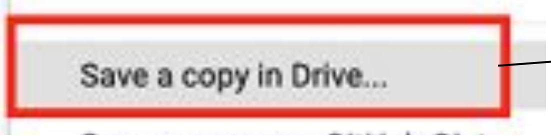


LAB03 APAI: DNN shrinking & quantization

Credits: *Davide Nadalini, Lorenzo Lamberti, Luca Bompani, Alberto Dequino, Francesco Conti. (University of Bologna)*

Contacts: lorenzo.lamberti@unibo.it, d.nadalini@unibo.it, alberto.dequino@unibo.it, luca.bompani5@unibo.it

4

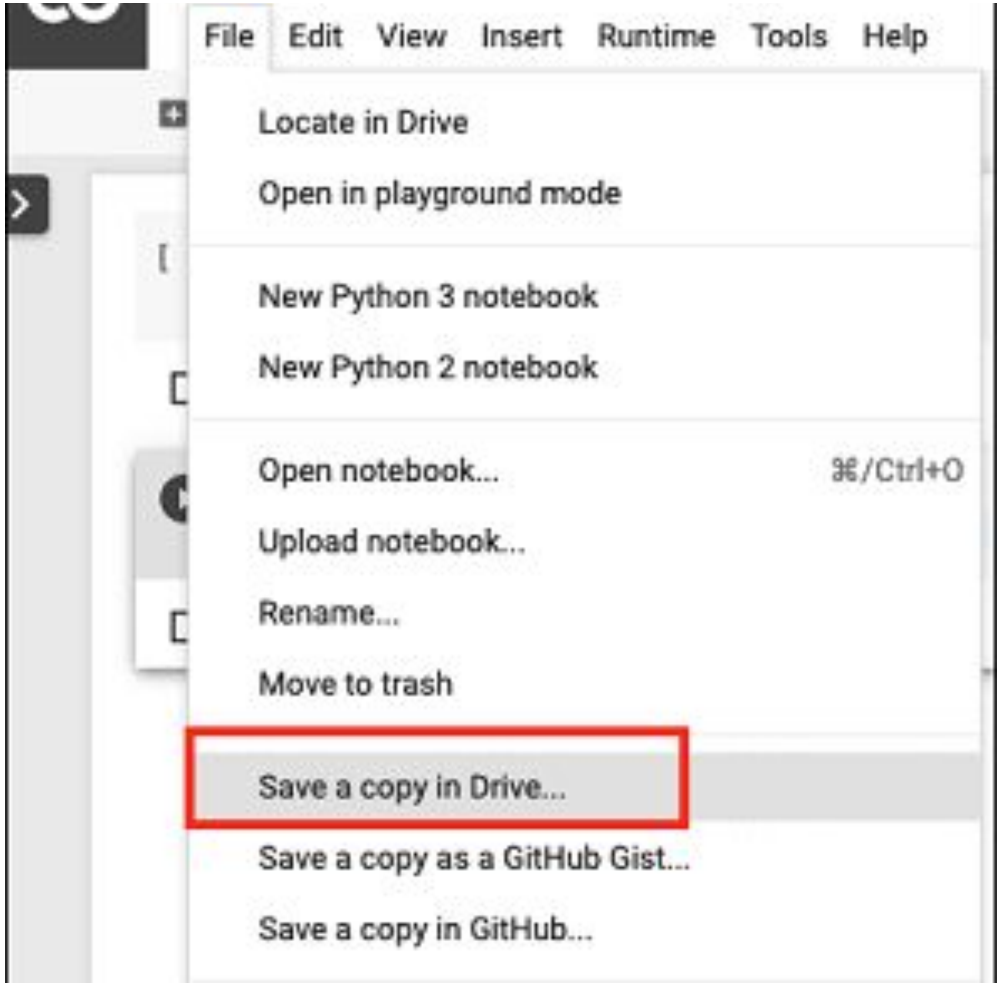


create your own copy in COLAB to modify it !

more details on step [4] of the setup

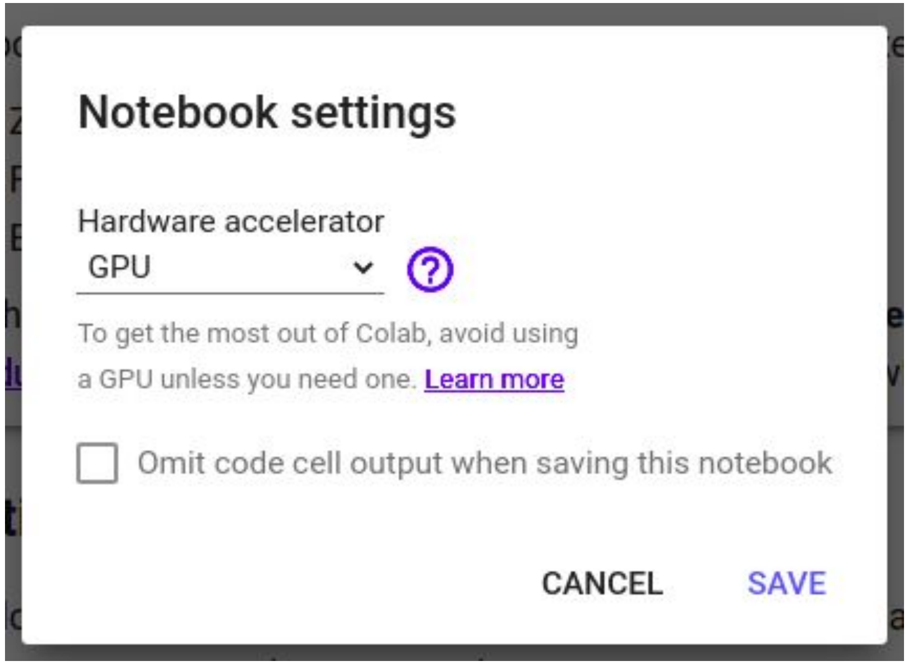
IMPORTANT:

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

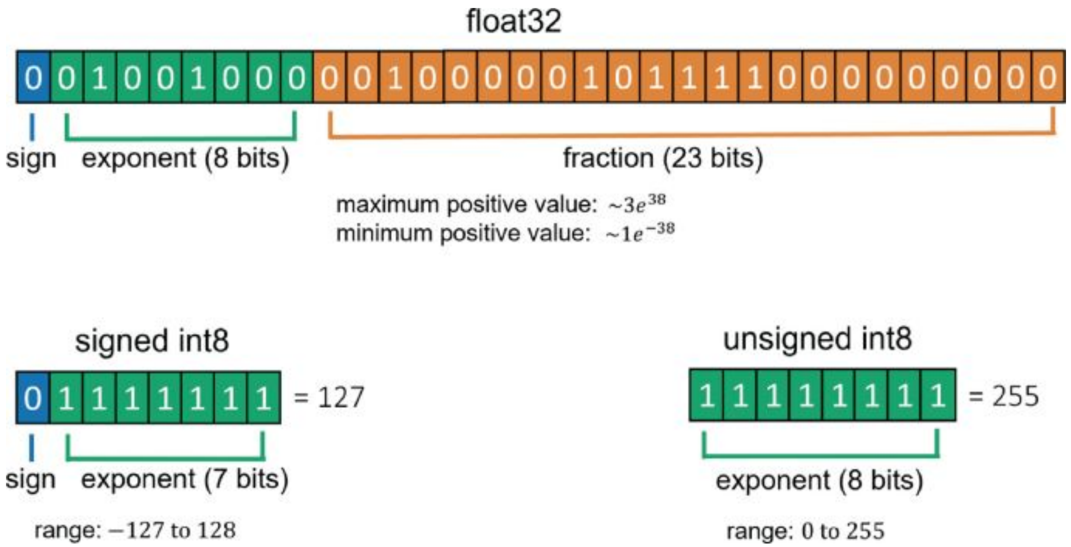
□ Omit code cell output when saving this notebook

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

A bit of theory

Data Types and Quantization

IEEE 754 32-bit Floating Point format



Training time: FP32 for numerical accuracy! However, 1 element of activation / weights = 4 bytes (memory hungry for inference!)

Deployment: quantization! Smaller number of bits for a single element (e.g., 8 bits = 1 byte), but less numerical precision!!

Remember Lab1? **Data stored in 8 bits** (instead of 16 or 32) easily **overflow** or cannot represent **values outside range**! One possible solution: **quantization-aware training**!

Sources: https://en.wikipedia.org/wiki/Single-precision_floating-point_format
https://link.springer.com/chapter/10.1007/978-3-031-24538-1_1

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

The LAB starts now !