Deep Neural Networks for Constrained Devices

Winter 2025

Teachers

Teacher	Email	Office
Christian Gianoglio	Christian.gianoglio@unige.it	DITEN, 3rd floor

General information

Aim of the course

The course provides the basic knowledge about the key aspects that should be considered when deploying a deep neural network on a constrained device. It offers a brief introduction to the topic and presents a survey of the most important tools and techniques. Practical examples of quantization, pruning, and the design of efficient architectures are provided using TensorFlow Lite. In addition, examples of deployment on STM32 microcontrollers are presented.

Teaching program

Introduction to the preliminary concepts, quantization and pruning, design of efficient architectures, use cases deployed on a STM32 device.

Language

Italian; if requested by foreign students, the course will be held in English.

Exam modality

Presentation of a project where the student exploits the techniques and the concepts learned during the course.

Bibliography

Notes by the teacher (in English), Tensorflow Lite documentation https://www.tensorflow.org/lite

Registration

Please, contact the teacher to register for the course. The location will be announced before the course begins. The course can also be attended through Microsoft Teams. Credential will be communicated after registration.

Timetable

Topic	Day	Time
Introduction	07/10/2025	09:30-12:30
Quantization	09/10/2025	09:30-12:30
Designing a Deep Network	14/10/2025	09:30-12:30
Use case	16/10/2025	09:30-12:30