Updated 07/08/2025

Sign up

ONNX Edge AI Deploy Training

2 days (14 hours)

Overview

ONNX Edge AI Deploy is a set of tools and practices for converting, optimizing and executing AI models in ONNX format on low-power devices such as embedded boards.

ONNX, for Open Neural Network Exchange, is an open standard that promotes interoperability between frameworks such as PyTorch and TensorFlow. This format facilitates the rapid, stable and optimized deployment of AI models on a wide range of Edge platforms.

Our ONNX Edge AI Deploy training course will enable you to master the complete cycle from preparation to execution of ONNX models in a constrained environment. You'll learn how to convert your models to ONNX format, optimize them for smooth operation on embedded hardware, and deploy them on platforms such as Jetson Nano, Raspberry Pi or Coral.

Emphasis is placed on the practical aspects of debugging, quantization and performance measurement to guarantee high-quality industrial deployment.

On completion of this course, you will be able to build a complete processing pipeline, from data to deployment, while ensuring maximum interoperability with market tools.

This course will introduce you to the latest stable version of ONNX Runtime v1.22.1 and the associated best practices.

Objectives

- Understand the architecture and role of the ONNX format
- Convert and optimize models for Edge targets
- Deploy models on embedded platforms

- Use ONNX Runtime and its hardware gas pedals
- Diagnose errors and monitor performance

Target audience

- Data / AI engineers
- Embedded systems developers
- Edge computing architects
- Researchers and makers wishing to industrialize an AI POC

Prerequisites

- Good command of Python and ML model manipulation
- Basic knowledge of application deployment (embedded Linux)
- Previous experience with PyTorch or TensorFlow recommended

ONNX Edge AI Deploy training program

Understanding ONNX and the Edge AI framework

- Introduction to ONNX: objectives, architecture, interoperability
- Overview of compatible frameworks: PyTorch, TensorFlow, etc.
- Key Edge AI concepts: hardware constraints and optimization
- Advantages of the ONNX format for cross-platform deployment
- Case studies: using ONNX models in production
- Workshop: Converting a PyTorch model to ONNX format

Exploring the ONNX format and tools

- Structure of an ONNX model file
- Using ONNX Checker to validate models
- Graphical visualization with Netron
- Debugging and inter-framework compatibility
- Current limitations of the ONNX format

Optimizing models for Edge environments

- Notions of dynamic and static quantization
- Dimension reduction: pruning and compression
- Using ONNX Runtime for the Edge
- Hardware acceleration: CPU, GPU, NPU
- Workshop: Optimizing an ONNX model with quantization and performance testing

Deploying ONNX Runtime on the Edge platform

- Overview of ONNX Runtime: versions, runtime options
- Configuration on Jetson Nano, Raspberry Pi, etc.
- Installation and integration in a Python project
- Benchmarks for latency and memory consumption
- Handling runtime errors

Creating a complete AI Edge pipeline

- Chaining: pre-processing ? ONNX model ? post-processing
- Use of real-time sensors and cameras
- Stand-alone or connected embedded execution
- Triggering conditional processing
- Workshop: Deploying a complete AI Edge application with ONNX Runtime

Monitor, update and industrialize

- Performance monitoring on Edge devices
- Deployment via Docker containers or automated scripts
- Update strategies for embedded models
- Security and robustness considerations
- Checklist for deploying Edge IA in production

Companies concerned

This training course is aimed at both individuals and companies, large or small, wishing to train their teams in a new advanced IT technology, or to acquire specific business knowledge or modern methods.

Positioning on entry to training

Positioning at the start of training complies with Qualiopi quality criteria. As soon as registration is finalized, the learner receives a self-assessment questionnaire which enables us to assess his or her estimated level of proficiency in different types of technology, as well as his or her expectations and personal objectives for the forthcoming course, within the limits imposed by the selected format. This questionnaire also enables us to anticipate any connection or security difficulties within the company (intra-company or virtual classroom) which could be problematic for the follow-up and smooth running of the training session.

Teaching methods

Practical course: 60% Practical, 40% Theory. Training material distributed in digital format to all participants.

Organization

The course alternates theoretical input from the trainer, supported by examples, with brainstorming sessions and group work.

Validation

At the end of the session, a multiple-choice questionnaire verifies the correct acquisition of skills.

Certification

A certificate will be awarded to each trainee who has completed the entire course.