Intro to ML-ops

Sveinung Myhre

ReLU NTNU

September 30, 2025

Outline

Introduction

Welcome to this brief introduction to ML-ops. Today we will cover:

- Docker containers
- Kubernetes
- GPU glossary
- Hands-on coding and pair-programming

Docker Containers

What are Docker Containers?

Lightweight, portable environments that package applications with their dependencies.

Key Benefits

- Consistency across environments
- Resource efficiency
- Easy deployment and scaling

Kubernetes

Container Orchestration

Kubernetes manages containerized applications across clusters of machines.

Key Features

- Automatic scaling
- Load balancing
- Self-healing
- Rolling updates

GPU Terminology for ML

- **CUDA**: Parallel computing platform for NVIDIA GPUs
- **Tensor Cores**: Specialized units for AI workloads
- **VRAM**: Video memory for storing model parameters
- **FP16/FP32**: Floating-point precision levels
- **Multi-GPU**: Using multiple GPUs for training

Pair Programming Session

Time for hands-on practice!

- Setting up ML containers
- Deploying with Kubernetes
- GPU optimization
- Best practices discussion

Thank you!
Questions?

sveinung.myhre@example.com