Deploying a Model for Inference at Production Scale (NVIDIA)

Duration: 08 hours

Learn how to scale your machine learning models to work effectively at production scale with hands-on exercises using NVIDIA Triton Inference Server and Prometheus.

Course Prerequisites:

Familiarity with at least one Machine Learning framework such as:

- PyTorch
- TensorFlow *
- ONNX
- TensorRT

About this Course

At scale machine learning models can interact with up to millions of users in a day. As usage grows, the cost of both money and engineering time can prevent models from reaching their full potential. It's these types of challenges that inspired creation of Machine Learning Operations (MLOps).

Learning Objectives

Practice Machine Learning Operations by:

- Deploying neural networks from a variety of frameworks onto a live Triton Server
- Measuring GPU usage and other metrics with Prometheus
- Sending asynchronous requests to maximize throughput

Upon completion, learners will be able to deploy their own machine learning models on a GPU server.

Topics Covered

- PyTorch
- Convolutional Neural Networks (CNNS)
- Data Augmentation
- Transfer Learning
- Natural Language Processing

Course Outline

- Getting Started
- Simple PyTorch Model
- HuggingFace Model
- Simple TensorFlow Model
- Simple TensorRT Model
- Advanced Inference
- Metrics