Model Parallelism: Building and Deploying Large Neural Networks

ID MPBDLNN **Price** 500.— €(excl. tax) **Duration** 1 day

Course Overview

Very large deep neural networks (DNNs), whether applied to natural language processing (e.g., GPT-3), computer vision (e.g., huge Vision Transformers), or speech AI (e.g., Wave2Vec 2) have certain properties that set them apart from their smaller counterparts. As DNNs become larger and are trained on progressively larger datasets, they can adapt to new tasks with just a handful of training examples, accelerating the route toward general artificial intelligence. Training models that contain tens to hundreds of billions of parameters on vast datasets isn't trivial and requires a unique combination of AI, high-performance computing (HPC), and systems knowledge.

Please note that once a booking has been confirmed, it is non-refundable. This means that after you have confirmed your seat for an event, it cannot be cancelled and no refund will be issued, regardless of attendance.

This course is part of the following Certifications

NVIDIA-Certified Professional: Generative AI LLMs (NCP-GENL)

Prerequisites

Familiarity with:

- Good understanding of PyTorch
- Good understanding of deep learning and data parallel training concepts
- Practice with deep learning and data parallel are useful, but optional

Course Objectives

In this workshop, participants will learn how to:

- Train neural networks across multiple servers
- Use techniques such as activation checkpointing, gradient accumulation, and various forms of model parallelism to overcome the challenges associated with large-model memory footprint
- Capture and understand training performance characteristics to optimize model architecture
- Deploy very large multi-GPU models to production using NVIDIA Triton™ Inference Server

Detailed Course Outline

Introduction

- Meet the instructor.
- Create an account at courses.nvidia.com/join

Introduction to Training of Large Models

- Learn about the motivation behind and key challenges of training large models.
- Get an overview of the basic techniques and tools needed for large-scale training.
- Get an introduction to distributed training and the Slurm job scheduler.
- Train a GPT model using data parallelism.
- Profile the training process and understand execution performance.

Model Parallelism: Advanced Topics

- Increase the model size using a range of memory-saving techniques.
- Get an introduction to tensor and pipeline parallelism.
- Go beyond natural language processing and get an introduction to DeepSpeed.
- Auto-tune model performance.
- Learn about mixture-of-experts models.

Inference of Large Models

- Understand the challenges of deployment associated with large models.
- Explore techniques for model reduction.
- Learn how to use TensorRT-LLM.
- Learn how to use Triton Inference Server.
- Understand the process of deploying GPT checkpoint to production.
- See an example of prompt engineering.

Final Review

- Review key learnings and answer questions.
- Complete the assessment and earn a certificate.
- Complete the workshop survey.

About Fast Lane

Fast Lane is a global, award-winning specialist in technology and business training as well as consulting services for digital transformation. As the only global partner of the three cloud hyperscalers- Microsoft, AWS and Google- and partner of 30 other leading IT vendors, Fast Lane offers qualification solutions and professional services that can be scaled as needed. More than 4,000 experienced Fast Lane professionals train and advise customers in organizations of all sizes in 90 countries worldwide in the areas of cloud, artificial intelligence, cyber security, software development, wireless and mobility, modern workplace, as well as management and leadership skills, IT and project management.

Fast Lane Services

- ✓ High End Technology Training
- ✓ Business & Soft Skill Training
- ✓ Consulting Services
- ✓ Managed Training Services
- ✓ Digital Learning Solutions
- ✓ Content Development
- ✓ Remote Labs
- ✓ Talent Programs
- ✓ Event Management Services

Training Methods

- ✓ Classroom Training
- ✓ Instructor-Led Online Training
- ✓ FLEX Classroom – Classroom & Online Hybrid
- ✓ Onsite & Customized Training
- ✓ E-Learning
- ✓ Blended & Hybrid Learning
- ✓ Mobile Learning

Technologies & Solutions

- ✓ Digital Transformation
- ✓ Artificial Intelligence
- ✓ Cloud
- ✓ Networking
- ✓ Cyber Security
- ✓ Wireless & Mobility
- ✓ Modern Workplace
- ✓ Data Center

Worldwide Presence
with high-end training centers around the globe

Multiple Awards
from vendors such as AWS, Microsoft, Cisco, Google, NetApp, VMware

Experienced SMEs
with over 19.000 combined certifications

Germany
Fast Lane Institute for Knowledge Transfer GmbH
Tel. +49 40 25334610
info@flane.de / www.flane.de

Austria
ITLS GmbH
(Partner of Fast Lane)
Tel. +43 1 6000 8800
info@itls.at / www.itls.at

Switzerland
Fast Lane Institute for Knowledge Transfer (Switzerland) AG
Tel. +41 44 8325080
info@flane.ch / www.flane.ch