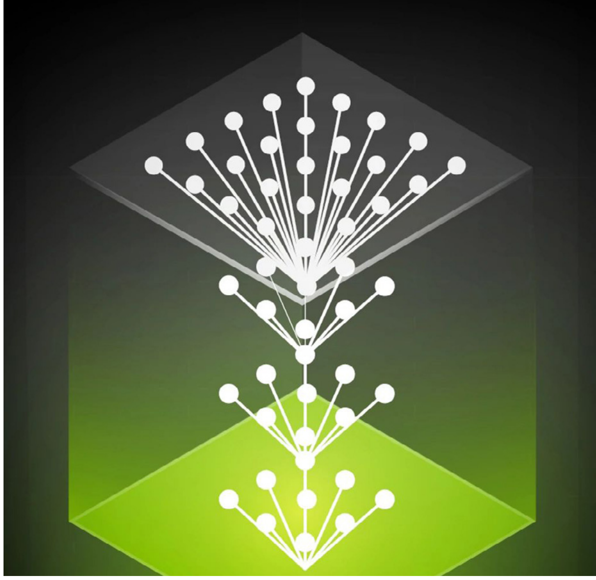
Datasheet

NVIDIA

Model Parallelism: Building and Deploying Large Neural Networks

NVIDIA Deep Learning Institute



Workshop Overview

Large language models (LLMs) and 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 LLMs and 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. The goal of this course is to demonstrate how to train the largest of neural networks and deploy them to production.

Learning Objectives

By participating in this workshop, you'll learn how to:

- Scale training and deployment of LLMs and neural networks across multiple nodes
- 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, multi-node models to production using NVIDIA TensorRT™-LLM

Overview	
Duration	8 hours
Price	**Contact us for pricing**.
Prerequisites	> Good understanding of **PyTorch** > Good understanding of **deep learning** and **data parallel** training concepts > Practice with **multi-GPU training** and **natural language processing** are useful, but optional
Tools, libraries, and frameworks	PyTorch, NVIDIA NeMo™ Framework, DeepSpeed, Slurm, TensorRT-LLM, NVIDIA Nsight™
Assessment type	Skills-based coding assessments evaluate learners' ability to train deep learning models on multiple GPUs.
Certificate	Upon successful completion of the assessment, participants will receive an NVIDIA DLI certificate to recognize their subject matter competency and support professional career growth.
Hardware Requirements	Desktop or laptop computer capable of running the latest version of Chrome or Firefox. Each participant will be provided with dedicated access to a fully configured, GPU-accelerated workstation in the cloud.
Language	English

Workshop Outline	
Introduction (15 minutes)	Meet the instructor. > Create an account at **courses.nvidia.com/join**
Introduction to Training of Large Models (120 minutes)	> 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.
Break (60 minutes)	
Model Parallelism: Advanced Topics (120 minutes)	> 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.
Break (15 minutes)	
Inference of Large Models (90 minutes)	> Understand the challenges of deployment associated with large models. > Explore techniques for model reduction. > Learn how to use TensorRT-LLM. > Understand the process of deploying GPT checkpoint to production. > See an example of prompt engineering.
Final Review (15 minutes)	> Review key learnings and answer questions. > Complete the assessment and earn a certificate. > Complete the workshop survey. > Learn how to set up your own AI application development environment
Next Steps	> **Building Conversational AI Applications** > **Data Parallelism: How to Train Deep Learning Models on Multiple GPUs** > **Building Transformer-Based Natural Language Processing Applications** > **Building RAG Agents for LLMs**

Why Choose NVIDIA Deep Learning Institute for Hands-On Training?

- Access workshops from anywhere with just your desktop/laptop and an internet connection. Each participant will have access to a fully configured, GPU-accelerated server in the cloud.
- Obtain hands-on experience with the most widely used, industry-standard software, tools, and frameworks.
- Learn to build deep learning and accelerated computing applications for industries, such as healthcare, robotics, manufacturing, accelerated computing, and more.
- Gain real-world expertise through content designed by NVIDIA and industry experts.
- Earn an NVIDIA DLI certificate to demonstrate your subject matter competency and support your career growth.

Ready to Get Started?

For the latest DLI workshops and trainings, visit **www.nvidia.com/dli**

For questions, contact us at **nvdli@nvidia.com**

© 2024 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, NeMo, Nsight, and TensorRT are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. All other trademarks and copyrights are the property of their respective owners. 3187848. MAR24