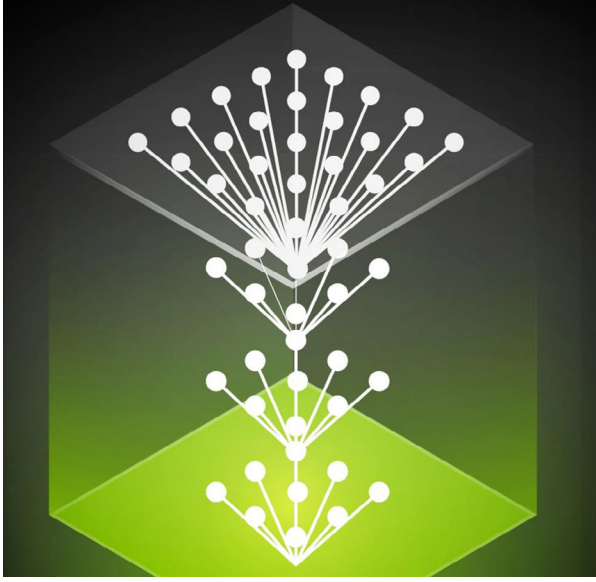
Datasheet

NVIDIA

模型并行 —— 构建和部署大型神经网络



课程概述

超大型的深度神经网络 (DNN)，无论是用于自然语言处理 (如 GPT)、计算机视觉 (如规模巨大的视觉 Transformer)，还是语音 AI (如 Wave2Vec 2)，都具有区别于其较小规模同类模型的特殊属性。基于海量数据集训练出来的 DNN 的规模越来越大，使得它们只需再经过少量示例训练就能够适应新任务要求，从而加快了向通用人工智能的迈进。使用庞大的数据集，训练包含数百亿到千亿参数的模型并非易事，这需要独特的方法来综合运用人工智能、高性能计算 (HPC) 和系统知识。本课程的目标就是学习如何训练超大型的神经网络并将其部署到生产中。

学习目标

完成本课程后，您将能够:

- 跨多个服务器训练神经网络
- 使用激活检查点、梯度累积和各种形式的模型并行等技术，来克服与大型模型内存占用相关的挑战
- 捕获并了解训练性能特征以优化模型架构
- 使用 NVIDIA® TensorRT™-LLM 将超大型多 GPU 模型部署到生产环境

课程概要	
课程时长	8 小时，课后 6 个月内可以继续访问课件（实验资源用量有限额）
课程模式	讲师实时授课，每位学员可使用云端完全配置的加速工作站实验练习
预备知识	> 熟悉 **PyTorch** > 熟悉**深度学习**和**数据并行**训练概念 > 先学习过**《数据并行 —— 用多 GPU 训练神经网络》**和**《构建基于 Transformer 的自然语言处理应用》**课程会很有帮助（可选）
使用的工具、库和框架	PyTorch, NeMo Framework, DeepSpeed, Slurm, TensorRT-LLM
课程测评问题类型	> 回答与课程内容相关的一系列问题 > 完成一个编程练习：要求将独立的 DNN 训练程序迁移到 DeepSpeed，将其执行分布到集群当中，并引入内存节省技术，这将允许进行有效的大规模训练。
培训证书	成功完成本课程和测试后，将获得 NVIDIA DLI 培训证书，证明在相关领域的能力，为职业发展提供证明。
课程语言	中文
学习此课程的硬件要求	您需要一台笔记本电脑或台式机，且能够运行最新版 Chrome 或 Firefox 浏览器。我们为您提供在云端完全配置的加速工作站的专用访问权限。
课程价格	> **AI 培训班**：每人 3500 元 (提供发票) > 企业专属培训：联系我们，微信添加 NVIDIALearn

为何选择 NVIDIA 深度学习培训中心 (DLI) 的实战培训?

- 学习 NVIDIA 与技术专家和行业领导者合作开发的课程，获取全球同步、技术领先和现实可用的专业开发技能和经验。
- 学习使用行业通用、标准的软件、工具和框架进行端到端的应用开发，能够在广泛的行业中构建基于深度学习、加速计算、图形与仿真和数据科学的应用。
- 系统化地学习理论，并使用云端完全配置的实验环境同步边练，高效提升实战开发能力。
- 获得 NVIDIA 全球开发者培训证书，加持专业培训认证，助力职业发展。

课程大纲	
课程介绍 (15 分钟)	> 讲师介绍 > 登录 DLI 学习平台
训练大模型 (120 分钟)	> 了解训练大型模型的需求和主要挑战 > 了解训练大规模所需的基本技术和工具 > 了解分布式训练和 Slurm 作业调度程序 > 使用数据并行训练 GPT 模型 > 分析训练过程并理解执行的性能
休息 (60 分钟)	
模型并行高级技能 (120 分钟)	> 使用一系列节省内存的技术来增加模型规模 > 了解 Tensor 和并行工作流 > 超越自然语言处理，了解 DeepSpeed > 自动调整模型性能 > 了解混合专家 (MoE) 模型
休息 (15 分钟)	
大模型推理 (120 分钟)	> 理解与大型模型相关的部署挑战 > 探索模型缩减技术 > 学习使用 NVIDIA TensorRT-LLM > 学习使用 NVIDIA Triton™ 推理服务器 > 理解将 GPT 检查点部署到生产环境的过程 > 查看提示工程的示例
总结 (30 分钟)	> 回顾今日所学 > 完成评估测试和获取证书 > 填写反馈表 > 了解如何设置您自己的 AI 应用开发环境
下一步	继续学习更多 DLI 课程： > **《数据并行 —— 用多 GPU 训练神经网络》** > **《构建基于 Transformer 的自然语言处理应用》** > **《构建大语言模型 RAG 智能体》**

准备好开始学习了吗?

查询更多 DLI 课程，请访问 **nvidia.cn/training**

如有疑问，请通过微信联系 DLI 小助手 (微信号 NVIDIALearn) 。