技术概述

NVIDIA GPU CLOUD 深度学习框架

NVIDIA GPU Cloud 优化框架容器指南

NVIDIA

简介

人工智能正在帮助解决人类所面临的某些疑难杂症：为传染病提供早期检测及寻找治疗方法、减少交通死亡事故、发现关键基础架构的缺陷以免它们带来安全隐患，等等。提高极限性能和管理底层技术恒定的变化速率是使用 AI 和深度学习的两大主要障碍。

NVIDIA 通过 NVIDIA GPU Cloud (NGC) 解决了这些问题。NGC 可为 AI 研究人员提供高性能的深度学习框架容器，为他们节省花在 IT 上的时间，使其有更多时间进行试验、发现有用见解并加快获得结果。

NVIDIA AI 深度学习

NVIDIA GPU Cloud 是一个针对深度学习优化的 GPU 加速云平台。NGC 管理着一个经过全面集成和优化的深度学习框架容器库，可充分利用 NVIDIA GPU。这些框架容器支持随时运行，包含所有必要的依赖项，例如 CUDA 运行时、NVIDIA 库以及操作系统。NVIDIA 对它们进行了调优、测试和验证，可在 Amazon EC2 P3 实例中 (即将推出其他云提供商) 使用 NVIDIA Volta™ 和 NVIDIA DGX 系统。NVIDIA 每月会更新这些容器，确保它们持续提供出色性能。

随时随地轻松启动和运行

NGC 容器注册通过 NVIDIA GPU Cloud 提供，可轻松利用新款 NVIDIA GPU 的强大功能。用户现在可以毫不费力地使用能够充分利用 NVIDIA GPU 强大功能的预集成容器创建深度神经网络 (DNN)。这让数据科学家、研究人员和工程师能够比以往更轻松地应对曾经认为无法克服的 AI 挑战，无论他们工作于配备齐全的实验室，还是利用云基础架构，都能游刃有余。

- **创造以分钟计算，而非以周计算** - 诸如 TensorFlow、PyTorch、MXNet 等的热门深度学习框架经过 NVIDIA 的严密调优、测试和验证，借助 NVIDIA Tesla V100 GPU 和 NVIDIA DGX 系统，可在 Amazon EC2 P3 实例中发挥十分出色的性能。这些框架是预先集成的易用容器，用户可以立即开展深度学习，降低时耗，并避免需手动进行的高难度软件集成。
- **跨平台深度学习** - 数据科学家和研究人员可在桌面、数据中心和云端中快速利用 NVIDIA GPU 构建、训练和部署深度神经网络模型。NGC 让他们能够灵活地在十分适宜的环境中开展工作，并根据需要即时提供可扩展性，从而帮助应对某些异常复杂的 AI 挑战。
- **持续更新** - NGC 提供的容器得益于 NVIDIA 的不断研发，可确保每个深度学习框架经过调优后，均能在全新 NVIDIA GPU 上实现很快的训练速度。NVIDIA 工程师不断优化库、驱动程序和容器，并且每月进行更新，以确保用户的深度学习投资逐渐获得更高回报。

NGC 容器注册

NGC 容器注册是一个 GPU 加速的深度学习软件库。其包含 CUDA 工具包、DIGITS 工作流，以及下列深度学习框架：NVCaffe、Caffe2、Microsoft Cognitive Toolkit (CNTK) 、MXNet、PyTorch、TensorFlow、Theano 和 Torch。

NGC 容器注册提供这些框架的容器化版本。这些框架，包括所有必要依赖项在内，构成了 NGC 深度学习堆栈。对于在建立定制深度学习解决方案时需要更多灵活性的用户，每个框架容器镜像还包含支持自定义修改和增强功能的框架源代码，以及完整的软件开发堆栈。

该平台软件的设计理念为尽量减少服务器上安装的操作系统和驱动程序，并通过 NVIDIA Docker Registry 在 NVIDIA Docker 容器内配置全部应用程序和 SDK 软件。图 1 展示了 NGC 深度学习堆栈层的图形布局。

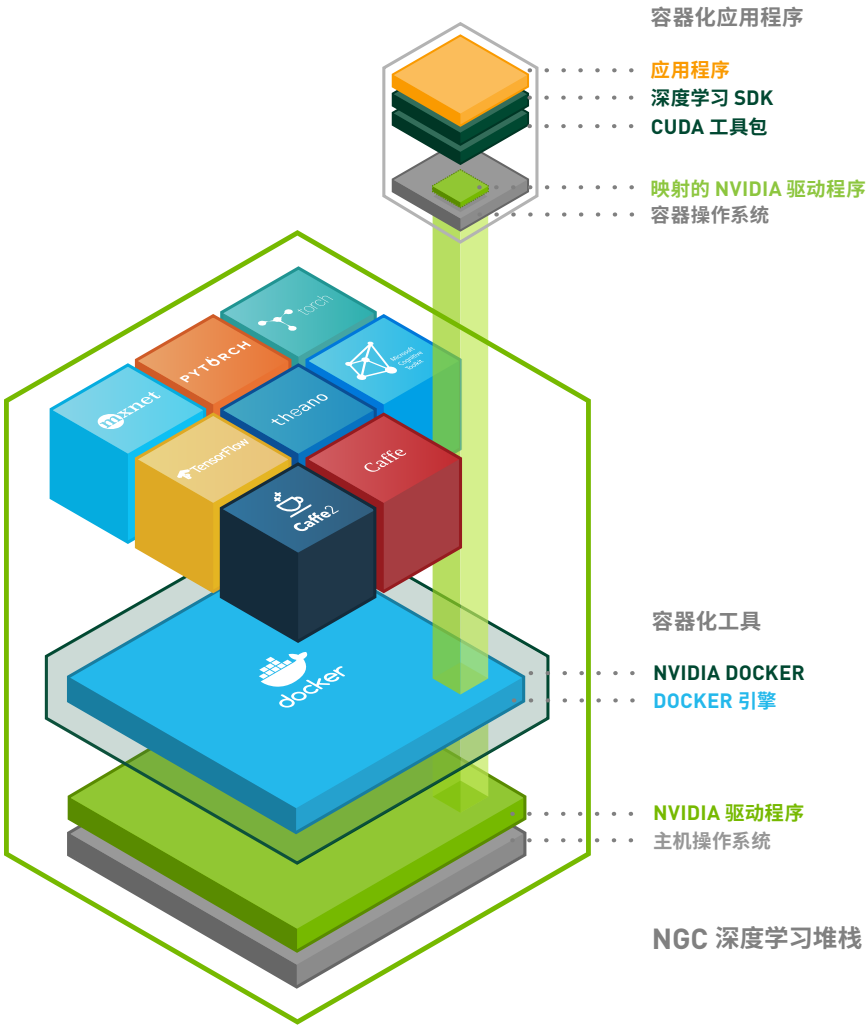


图 1：NVIDIA Docker 会在启动时将 NVIDIA 驱动程序和 GPU 的用户模式组件加载到 Docker 容器中。

为实现利用 GPU 的 Docker 镜像的可移植性，NVIDIA 开发了开源项目 NVIDIA Docker，该项目提供一个命令行工具，用于在启动时就将 NVIDIA 驱动程序的用户模式组件和 GPU 全部加载到 Docker 容器中。nv-docker 本质上是一个围绕 Docker 的包装器，能够以透明的方式为容器配置必要的组件，以在 GPU 上执行代码。Docker 容器提供了一个机制，用于将 Linux 应用程序与其所有的库、配置文件及环境变量捆绑在一起，这样不论在什么 Linux 系统上运行或是在相同主机实例间，其执行环境始终相同。Docker 容器为仅用户模式，所有从容器调用的内核都由主机系统内核处理。

分层方法

深度学习框架是软件堆栈的一部分，而软件堆栈由多个层组成。每个层依赖堆栈中位于其下方的层。该软件架构具有很多优势：

- 每个深度学习框架都位于单独的容器内，所以每个框架都能使用不同版本的库，比如 libc、cuDNN 等，并且不会相互影响。
- 分层容器的一个主要原因是，用户可以锁定所需要的目标体验。
- 为提高性能或修复问题，深度学习框架经过多次改进，现在注册中已有新版本容器。
- 系统易于维护，且由于应用程序并非直接安装于操作系统上，所以操作系统镜像非常干净。
- 可无缝提供安全更新、驱动程序更新及操作系统补丁。

为何使用框架？

创建框架是为了更轻松、高效地研究和应用深度学习。使用框架的主要好处包括：

- 框架可专门为深度神经网络（DNN）训练所需的计算提供支持高度优化 GPU 的代码。
- NVIDIA 的框架经过调优和测试，可提供非常出色的 GPU 性能。
- 借助这些框架，用户可以通过简单的命令行或 Python 等脚本语言接口访问代码。
- 许多功能强大的 DNN 都可通过这些框架来训练和部署，而无需编写任何 GPU 或复杂的编译代码，与此同时仍从 GPU 加速带来的训练速度提高中受益。

NGC 深度学习堆栈容器

Caffe

NVCAFFE

Caffe7 深度学习框架在设计时已将灵活性、速度和模块化考虑在内。该框架最初是由伯克利视觉和学习中心 (BVLC) 及社区贡献者共同开发。

NVCaffe 是由 NVIDIA 维护的 BVLC Caffe 分支版本，专门针对 NVIDIA GPU（尤其是多 GPU 配置）进行调优。NVCaffe 包括：

- 混合精度支持。允许以 64、32 或 16 位格式存储和/或计算数据。用户可在每个层上定义精度（向前阶段和向后传输阶段可以不同），也可以为整个网络设置默认精度。
- 集成 cuDNN v6。
- 自动选择最佳 cuDNN 卷积算法。
- 集成 v1.3.4 的 NCCL 库，可提升多 GPU 扩展能力。
- 为数据和参数存储、I/O 缓冲和卷积层工作空间优化了 GPU 内存管理。
- 并行架构数据解析器和转换器，可提升 I/O 性能。
- 在多 GPU 系统上并行架构向后传播和梯度简化。
- 通过融合的 CUDA 内核快速实施求解器，实现权重和历史记录更新。
- 可为跨多个 GPU 的内存负载应用多 GPU 测试阶段。
- 向后兼容 BVLC Caffe 和 NVCaffe 0.15。
- 优化模型扩展集合（包括 16 位浮点示例）。

Caffe2

CAFFE2

Caffe2 是一种深度学习框架，可在友好的基于 Python 的 API 中轻松表示所有模型类型，例如卷积神经网络 (CNN)、递归神经网络 (RNN) 等，并能够使用十分高效的 C++ 和 CUDA 后端进行执行。

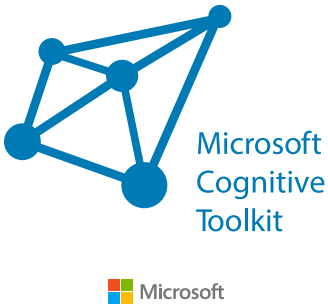
用户借助它能够非常灵活地结合使用高级运算和表达式运算来构架其模型以进行推理或训练，然后通过相同 Python 接口运行，轻松实现可视化，或序列化所创建的模型以及直接使用底层 C++ 实现。

Caffe2 支持单一和多 GPU 执行以及多节点执行。

以下列表总结了 NGC 深度学习堆栈 Caffe2 的优化和变更：

- 采用新版的 cuDNN。
- 微调性能。

- GPU 加速镜像输入管道。
- 自动选择最佳卷积算法。



MICROSOFT COGNITIVE TOOLKIT

Microsoft Cognitive Toolkit (CNTK) 是统一的深度学习工具包，让用户可轻松实现并合并前馈深度神经网络 (DNN)、CNN 和 RNN 等热门模型类型。

Microsoft Cognitive Toolkit 跨多个 GPU 和服务器进行自动微分和并行化，以执行随机梯度下降 (SGD) 学习。Microsoft Cognitive Toolkit 可称为 Python 或 C++ 应用程序库，或是作为使用 BrainScript 模型描述语言的独立工具来执行。

以下列表总结了 NGC 深度学习堆栈 Microsoft Cognitive Toolkit 的优化和变更：

- 采用新版的 cuDNN。
- 集成支持 NVLink 的最新版 NCCL，可提升多 GPU 扩展性。使用数据并行 SGD 时，支持 NVLink 的 NCCL 可将 ResNet-50 训练性能提升 2 倍。
- 改进图像阅读器管道，允许 AlexNet 以超过 12,000 张图片/秒的速度进行训练。
- 为多 GPU 训练的每块 GPU 减少高达 2 GB 的 GPU 显存用度。
- 扩大卷积支持。
- 优化可减少 cuDNN 工作空间所需的显存占用。

MXNET

MXNet 是专为提高效率和灵活性而设计的深度学习框架，让您可以混合符号编程和命令编程，以最大限度提升效率和生产力。

MXNet 的核心是动态依赖调度程序，可自动即时并行处理符号式操作和命令式操作。在调度程序之上增加图形优化层可加快符号执行速度并提高内存效率。MXNet 轻量便携，并且可扩展至多个 GPU 和多台机器上。

以下列表总结了 NGC 深度学习堆栈 MXNet 的优化和变更：

- 采用新版的 cuDNN。
- 改进输入管道，便于处理图像。
- 优化嵌入层 CUDA 内核。
- 优化张量传播和简化 CUDA 内核。

PYTORCH

PyTorch 是可提供两种高级功能的 Python 软件包：

- 拥有强大 GPU 加速功能的张量计算（如 numpy）。
- 构建于基于磁带的自动调整系统的深度神经网络。

您可以重新使用喜欢的 Python 软件包（如 numpy、scipy 和 Cython）来根据需要扩展 PyTorch。

以下列表总结了 NGC 深度学习堆栈 PyTorch 的优化和变更：

- 采用新版的 cuDNN。
- 集成支持 NVLink 的全新版 NCCL。
- 缓冲待 NCCL 通信的参数，以降低延迟用度。
- 扩大卷积支持。
- 用于避免不必要的数据副本和缓冲归零的优化。

TENSORFLOW

TensorFlow 是使用数据流图表进行数值计算的开源软件库。图像中的节点代表数学运算，而图像边缘则代表节点间流动的多维数据阵列（张量）。您可通过此灵活的架构，将计算部署到桌面、服务器或移动设备中的一个或多个 CPU 或 GPU，而无需重写代码。

TensorFlow 最初由 Google Brain 团队（隶属于 Google 机器智能研究组织）的研究人员和工程师开发，旨在进行机器学习和深度神经网络研究。该系统具有极佳的通用性，也适用于其他各种领域。

为了可视化 TensorFlow 结果，TensorFlow Docker 镜像也包含 TensorBoard。TensorBoard 是一套可视化工具。例如，您可以查看训练记录以及模型外观。

以下列表总结了 NGC 深度学习堆栈 TensorFlow 的优化和变更：

- 采用新版的 cuDNN。
- 集成支持 NVLink 的最新版 NCCL，可提升多 GPU 扩展性。使用数据并行 SGD 时，支持 NVLink 的 NCCL 可将 ResNet-50 训练性能提升 2 倍。
- 默认支持融合颜色调整内核。
- 默认支持使用非融合的 Winograd 卷积算法。

THEANO

Theano 是一个 Python 库，能让您有效定义、优化和评估涉及多维阵列的数学表达式。自 2007 年起，Theano 便一直为大规模计算密集型科学调查提供支持。

以下列表总结了 NGC 深度学习堆栈 Theano 的优化和变更：

- 采用新版的 cuDNN。
- 运行时代码生成：更快速地评估表达式。
- 广泛的单元测试和自我验证：检测与诊断多种类型的错误。

TORCH

Torch 是一种科学计算框架，可为深度学习算法提供广泛支持。由于简单且快速的脚本语言 Lua 以及底层 C/CUDA 实现，Torch 易于使用且效率很高。

Torch 提供热门的神经网络和优化库，不但容易使用，而且还提供超大灵活性，以构建复杂的神经网络拓扑。

以下列表总结了 NGC 深度学习堆栈 Torch 的优化和变更：

- 采用新版的 cuDNN。
- 集成最新版、采用 NVLink 支持的 NCCL，以提升多 GPU 扩展性。使用数据并行 SGD 时，支持 NVLink 的 NCCL 可将 ResNet-50 训练性能提升 2 倍。
- 缓冲待 NCCL 通信的参数，以降低延迟用度。
- 针对递归网络（RNN、GRU、LSTM）的 cuDNN 绑定（包括持久性版本），可大幅提升小批量训练的性能。
- 扩大卷积支持。
- 支持向 cuDNN 例程的 16 和 32 位浮点（FP16 和 FP32）数据输入。
- 支持对 FP16 张量的运算（使用 FP32 算法）。

DIGITS

NVIDIA 深度学习 GPU 训练系统（DIGITS）将深度学习的强大功能交由工程师和数据科学家掌控。

DIGITS 并不是框架。DIGITS 是 Caffe 和 Torch 的包装器；会提供连接这些框架的图形网络接口，而非直接在命令行上处理它们。

DIGITS 可用于快速训练高度准确的深度神经网络（DNN），以执行图像分类、分割以及物体检测等任务。DIGITS 可简化常见的深度学习任务，例如管理数据、在多 GPU 系统上设计和训练神经网络、使用高级可视化技术实时监控性能，以及从结果浏览器中选择性能更好的模型用于部署。DIGITS 采用完全交互式的设计，因此数据科学家可以专注于设计和训练网络，而不必进行编程和调试。

NVIDIA GPU Cloud 助力 AI 蓬勃发展

NVIDIA GPU Cloud 拥有一个经集成与优化的综合深度学习软件库。NVIDIA 利用其多年来在 AI 领域的丰富研发经验，在 NGC 容器注册中为每位用户提供可随时运行的高性能软件，并将其为深度学习框架带来的增强功能贡献给开源社区。

NVIDIA 在构建新版本的框架、驱动程序和硬件后，持续进行改进和更新，以确保每个部分以良好的状态共同发挥作用，并提供卓越性能，为用户有效减少测试和集成方面长久存在的重负。NGC 容器注册提供的框架让数据科学家和研究人员几乎在每个学科和行业均实现了重大的深度学习突破，并帮助他们解决当今 AI 领域的某些艰巨挑战。

如需了解有关 NGC 的更多信息并观看入门视频，请访问：
www.NVIDIA.cn/cloud

如需注册 NGC，请访问：
www.NVIDIA.cn/ngcsignup

NVIDIA