技術概要

NVIDIA GPU CLOUD ディープラーニング ソフトウェア

最適化されたディープラーニング コンテナーのガイド

NVIDIA

はじめに

AI は、感染症の早期発見と治療法の探索、交通事故死の削減、事故発生前の重大なインフラ欠陥の発見など、人類が直面する複雑な 問題を解決するのに役立てられています。AI とディープラーニング利用における 2 つの大きな課題は、パフォーマンスの最大化と、絶え間なく変化する基盤技術の管理です。

これを解決できるのが NVIDIA GPU Cloud (NGC) です。NGC では、パフォーマンスを重視して設計されたディープラーニング ソフトウェア コンテナーを活用して作業効率を高めることができます。AI 研究者は、IT に費やす時間を短縮し、より多くの時間を実験、考察、成果達成に充てられるようになります。

NVIDIA AI によるディープラーニング

NVIDIA GPU Cloud は、ディープラーニング向けに最適化された GPU 活用クラウド プラットフォームであり、NVIDIA GPU を最大限に活用することができるディープラーニング ソフトウェアの総合カタログです。コンテナーには、NVIDIA® CUDA® Toolkit、NVIDIA ディープラーニング ライブラリやオペレーティング システムなど必要なすべての要素が含まれており、すぐに実行することができます。また、コンテナーは NVIDIA DGX™ Systems、NVIDIA TITAN (NVIDIA Volta と NVIDIA Pascal™ を搭載)、NVIDIA Quadro® GV100、GP100、P6000、および Amazon EC2、Google Cloud Platform、Oracle Cloud Infrastructure 等の対応パブリック クラウド プロバイダーにおける動作が調整、テスト、認定されています。NVIDIA では最高の性能を維持できるように、コンテナー イメージを毎月更新しています。

いつでもどこでもすぐに運用開始

NVIDIA GPU Cloud の NGC コンテナー レジストリでは、最新 NVIDIA GPU のパワーを簡単に活用することができます。ユーザーは、NVIDIA GPU を最大限に活かした統合済みの高性能コンテナーを使用して、ディープ ニューラル ネットワーク (DNN) をすばやく作成することができます。データ サイエンティスト、研究者、技術者は、デスクトップ、完全に整備されたラボ、クラウド インフラストラクチャなどを組み合わせたさまざまな環境で AI を活用することで、これまで不可能とされていた課題にも挑戦することができます。

- **わずか数分でイノベーションを実現** - TensorFlow、PyTorch、MXNet、NVIDIA TensorRT™ をはじめとする優れたディープラーニング ソフトウェアは、NVIDIA DGX Systems、NVIDIA TITAN (NVIDIA Volta と

NVIDIA Pascal を搭載)、NVIDIA Quadro GV100、GP100、P6000、および対応パブリック クラウド プロバイダー (Amazon EC2、Google Cloud Platform、Oracle Cloud Infrastructure) の Volta/Pascal GPU インスタンスで、最大限のパフォーマンスを発揮するように調整、テスト、認定されています。ソフトウェアは、すぐにディープラーニング作業を開始できるように使いやすい統合済みコンテナーで提供されるため、ユーザーは時間のかかる複雑な統合作業を実行する必要がありません。

- **さまざまなプラットフォームでディープラーニング** - データ サイエンティストと研究者は、デスクトップ、データセンター、クラウドの NVIDIA GPU でディープ ニューラル ネットワーク モデルを迅速に構築、トレーニング、展開することができます。NGC は、最適な作業環境を作り出す柔軟性と、必要に応じて即座に対応できるスケーラビリティによって、非常に複雑な AI の課題の解決を支援します。
- **常に最新** - NGC で使用可能なディープラーニング コンテナーには、NVIDIA の継続的な開発の成果が反映されます。NVIDIA の技術者は、ライブラリ、ドライバー、コンテナーを毎月更新し、ユーザーのディープラーニングへの投資効果をさらに高められるよう、最適化を行っています。

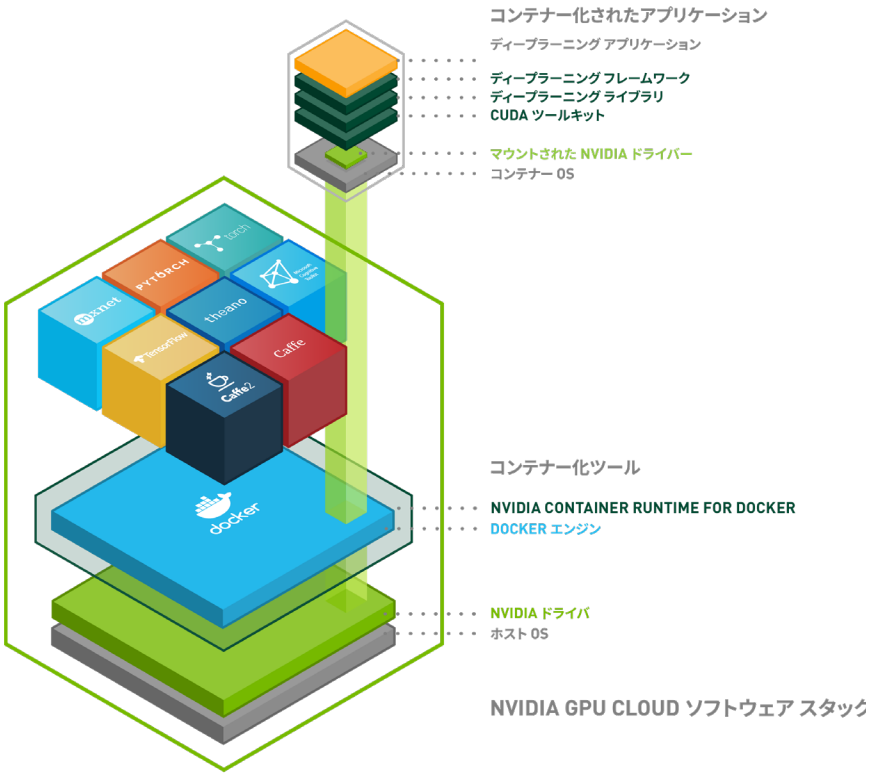
NGC コンテナー レジストリ

NGC コンテナー レジストリは、GPU アクセラレーション ディープラーニング ソフトウェアのカタログです。これには、NVIDIA CUDA Toolkit、NVIDIA DIGITS™、NVIDIA TensorRT、および NVCaffe、Caffe2、Microsoft Cognitive Toolkit (CNTK)、MXNet、PyTorch、TensorFlow、Theano、Torch の各ディープラーニング フレームワークが含まれます。

NGC コンテナー レジストリは、すべての依存要素を含むソフトウェアのコンテナー化されたバージョンを提供します。コンテナー内の最適化されたソフトウェア セットは、NVIDIA GPU Cloud ソフトウェア スタックと呼ばれます。カスタムのディープラーニング ソリューションを柔軟に構築したいユーザー向けに、各フレームワークのコンテナー イメージには、完全なソフトウェア開発スタックに加えて、変更や拡張をカスタマイズできるフレームワーク ソース コードが含まれています。

プラットフォームは、サーバーにインストールされている最小限の OS とドライバーを中心に設計されており、コンテナー内のすべてのアプリケーションとソフトウェア開発キット (SDK) ソフトウェアはレジストリを介してプロビジョニングされます。図 1 は、NGC Software Stack の層のレイアウトです。

図 1: NVIDIA Container Runtime for Docker は、起動時に NVIDIA ドライバーのユーザー モード コンポーネントと GPU を Docker コンテナーにマウントします。



GPU を活用した Docker イメージをポータブルにするために、NVIDIA はオープンソース プロジェクトとして NVIDIA Container Runtime for Docker を開発しました。これは、起動時に NVIDIA ドライバーのユーザー モード コンポーネントと GPU を Docker コンテナーにマウントするコマンドライン ツールです。nv-docker は、GPU でコードを実行するコンポーネントをコンテナーと共に透過的にプロビジョニングするためのラッパーです。Docker コンテナーとは、Linux アプリケーションが Linux システム上や同じホストのインスタンス上での実行環境を統一するメカニズムで、アプリケーションとライブラリ、構成ファイル、環境変数などがバンドルされています。Docker コンテナーはユーザー モードに限定されており、コンテナーからのすべてのカーネル呼び出しは、ホスト システムのカーネルによって処理されます。

階層型アプローチ

ディープラーニング フレームワークは、複数の層で構成されるソフトウェア スタックの一部です。各層は、スタック内の下にある層に依存します。このソフトウェア アーキテクチャには、次のような多くのメリットがあります。

- 各ディープラーニング フレームワークまたはアプリケーションが個別のコンテナーに格納されるため、互いに干渉することなく libc、cuDNN などさまざまなバージョンのライブラリを使用することができます。

- 層別コンテナーでは、ユーザーが必要とするエクスペリエンスを実現することができます。
- ディープラーニング フレームワークとアプリケーションに関するパフォーマンス改善やバグ修正のたびに、コンテナーの新しいバージョンがレジストリで提供されます。
- システムのメンテナンスが容易であるほか、フレームワークまたはアプリケーションが OS に直接インストールされないため、OS のイメージをクリーンに保つことができます。
- セキュリティ更新、ドライバー更新、OS の修正プログラムがシームレスに提供されます。

フレームワークを使用するメリット

フレームワークは、ディープラーニングの研究と応用のアクセス性と効率を高めることを目的としています。フレームワークを使用する主なメリットは次のとおりです。

- フレームワークで提供される高度に最適化された GPU 対応コードは、ディープ ニューラル ネットワーク (DNN) のトレーニングの演算処理に特化しています。
- NVIDIA のフレームワークは、最大限の GPU パフォーマンスを引き出せるように調整およびテストされています。
- 簡単なコマンドラインや Python などのスクリプト言語インターフェイスを使用してコードにアクセスすることができます。
- GPU 用コードや複雑なコンパイル済みコードを作成しなくても、複数の強力な DNN をトレーニングおよび実装できるだけでなく、GPU アクセラレーションによってトレーニングを高速化することもできます。

NGC ディープラーニングのコンテナー

このセクションでは、NGC で使用可能なディープラーニング ソフトウェアのコンテナーについて説明します。各コンテナーは、最新の NVIDIA ディープラーニング ライブラリの cuDNN、cuBLAS、NCCL との統合などを含め、毎月更新されます。

NVIDIA CUDA Deep Neural Network ライブラリ (cuDNN) は、ディープ ニューラル ネットワークのプリミティブの GPU アクセラレーション ライブラリです。cuDNN には、畳み込み、プーリング、正規化、アクティベーション層などの高度に調整された標準ルーチンが実装されています。

NVIDIA cuBLAS ライブラリは、標準の基本線形代数サブルーチン (BLAS) の GPU アクセラレーション実装です。cuBLAS API では、計算量が膨大な操作を、単一 GPU で処理するかマルチ GPU 構成に効率的に分散するかして、アプリケーションを高速化することができます。

NVIDIA Collective Communications Library (NCCL) では、NVIDIA GPU 向けにパフォーマンスが最適化されたマルチ GPU およびマルチノードの集合通信プリミティブを実装しています。NCCL で提供される all-gather、all-reduce、broadcast、reduce、reduce-scatter などのルーチンは、PCIe と NVLink の高速相互接続によって高帯域幅を達成するように最適化されています。

Caffe

NVCAFFE

Caffe は、柔軟性、速度、およびモジュール性を念頭に置いて作成されたディープラーニング フレームワークで、元は Berkeley Vision and Learning Center (BVLC) とコミュニティ参加者によって開発されたものです。

NVCaffe は、NVIDIA が管理する BVLC Caffe のフォークで、NVIDIA GPU (特にマルチ GPU 構成) 向けに調整されています。最新の機能強化については、「**NVCaffe コンテナーのリリース ノート (英語)**」を参照してください。

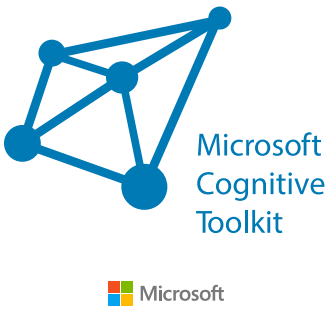
Caffe2

CAFFE2

Caffe2 は、畳み込みニューラル ネットワーク (CNN) やリカレント ニューラル ネットワーク (RNN) などの任意のモデル タイプを、使いやすい Python ベースのアプリケーション プログラミング インターフェイス (API) で簡単に表現し、効率的な C++ と CUDA バックエンドで実行するためのディープラーニング フレームワークです。

表現力の高いハイレベルな操作により、推論、学習を問わず、柔軟にモデルを構築することができます。こうして構築されたモデルは、同じ Python インターフェイスを用いて容易に可視化したり、あるいは作成されたモデルをシリアライズして、コアの C++ 実装を直接利用することもできます。Caffe2 は、シングル GPU、マルチ GPU、マルチノードの実行をサポートします。

最新の機能強化については、「**Caffe2 コンテナーのリリース ノート (英語)**」を参照してください。



Microsoft

MICROSOFT COGNITIVE TOOLKIT

Microsoft Cognitive Toolkit (旧称 CNTK) は、フィードフォワード ディープ ニューラル ネットワーク (DNN)、CNN、RNN などの一般的なモデル タイプを簡単に実現したり組み合わせたりすることができる、統合型ディープラーニング ツールキットです。

Microsoft Cognitive Toolkit では、自動微分を使った確率的勾配降下法 (SGD: Stochastic Gradient Descent)による学習を、複数の GPU 及びサーバーにわたる並列処理を活用して実装することができます。Python アプリケーションまたは C++ アプリケーションからライブラリとして呼び出すか、または BrainScript モデルの記述言語を使用してスタンドアロン ツールとして実行することができます。

最新の機能強化については、「**Microsoft Cognitive Toolkit コンテナーのリリース ノート (英語)**」を参照してください。

MXNET

MXNet は、効率と柔軟性の両方を考慮して設計されたディープラーニング フレームワークであり、記号プログラミングと命令型プログラミングの組み合わせにより、効率と生産性を最大化することができます。

MXNet は、記号と命令の両方の演算を自動的に並列化して迅速に処理する動的な依存要素スケジューラを基盤としています。スケジューラの上に構築されるグラフ最適化層が記号演算を高速化し、メモリを効率化します。また、移植性が高く軽量なため、複数の GPU やマシンにスケーリングできます。

最新の機能強化については、「**MXNet コンテナーのリリース ノート (英語)**」を参照してください。

PYTORCH

PyTorch は、次の 2 つのハイレベルな機能を提供する Python パッケージです。

- 強力な GPU アクセラレーションによるテンソル計算 (numpy など)
- テープベースの Autograd システムに基づいたディープ ニューラル ネットワーク

必要に応じて、numpy、scipy、Cython などの使い慣れた Python パッケージを再使用して PyTorch を拡張することもできます。

最新の機能強化については、「**PyTorch コンテナーのリリース ノート (英語)**」を参照してください。

TENSORFLOW

TensorFlow は、データ フロー グラフを使用する数値演算のためのオープン ソース ソフトウェア ライブラリです。グラフのノードは数学的演算を表し、グラフのエッジはその間を流れる多次元データ配列 (テンソル) を表します。この柔軟なアーキテクチャにより、コードの修正なしに、デスクトップ、サーバー、またはモバイル デバイスの 1 つ以上の CPU または GPU に演算処理を展開することができます。

TensorFlow は当初、機械学習とディープ ニューラル ネットワークの研究を行うために、Google の Machine Intelligence 研究組織内の Google Brain チームの研究者と技術者によって開発されました。現在では、他のさまざまなドメインにも適用できるまで十分に一般化されています。

イメージには、TensorFlow の結果を可視化するツール TensorBoard も付属しています。これにより、トレーニング履歴やモデルの外観などを表示することができます。

最新の機能強化については、「**TensorFlow コンテナーのリリース ノート (英語)**」を参照してください。

THEANO

Theano は、多次元配列を含む数式を効率的に定義、最適化、および評価するための Python ライブラリです。Theano は、2007 年から大規模な演算を行う科学調査の分野を牽引しています。

最新の機能強化については、「**Theano コンテナーのリリース ノート (英語)**」を参照してください。

TORCH

Torch は、幅広いディープラーニング アルゴリズムをサポートした科学計算フレームワークです。Lua という簡単で高速なスクリプト言語と C/CUDA 基盤を採用しており、非常に使いやすく効率的です。

Torch に含まれるニューラル ネットワークと最適化の一般的なライブラリは、簡単に使用できる一方で、複雑なニューラルネットワーク トポロジの構築にもきわめて柔軟に対応することができます。

最新の機能強化については、「**Torch コンテナーのリリース ノート (英語)**」を

参照してください。

DIGITS

DIGITS

NVIDIA Deep Learning GPU Training System (DIGITS) は、技術者とデータサイエンティスト向けに高いディープラーニング機能を提供します。

DIGITS はフレームワークではありません。Caffe と Torch フレームワークをコマンドラインで直接処理する代わりに、グラフィカルな Web インターフェイスを提供するラッパーです。

DIGITS を使用すると、画像分類、セグメンテーション、物体検出の各タスクで、高精度なディープ ニューラル ネットワーク (DNN) をすばやくトレーニングすることができます。また、データ管理、マルチ GPU システムにおけるニューラル ネットワークの設計とトレーニング、高度な可視化によるパフォーマンスのリアルタイム監視に加えて、展開用の最良モデルを結果ブラウザーから選択するなど、一般的なディープラーニング タスクを簡素化することができます。DIGITS は完全にインタラクティブでプログラミングやデバッグが不要であるため、ユーザーはネットワークの設計とトレーニングに専念することができます。

最新の機能強化については、「**DIGITS コンテナーのリリース ノート (英語)**」を参照してください。

TensorRT

TENSORRT

NVIDIA TensorRT は、NVIDIA GPU での高性能推論を可能にする C++ ライブラリです。テンソルと層のマージ、重み変換、効率的な中間データ形式の選択、層パラメーターや測定結果に基づく大きなカーネル　カタログからの抽出などにより、取得したネットワーク定義を最適化します。

TensorRT には、最適化オプションとして、高速で精度を抑えた Pascal GPU と Volta GPU の機能を利用できるインフラストラクチャが含まれます。

TensorRT 開発用に提供されている使いやすいコンテナーでは、TensorRT サンプルの構築、変更、実行などを行うことができます。詳細については、 **NVIDIA Deep Learning SDK のドキュメント** (英語) を参照してください。

最新の機能強化については、「**TensorRT コンテナーのリリース ノート (英語)**」を参照してください。

NVIDIA GPU Cloud で AI を高速化

NVIDIA GPU Cloud は、統合および最適化されたディープラーニング ソフトウェアの包括的なカタログを備えています。NVIDIA は AI における長年の研究開発の成果を活かして、ユーザー向けに NGC コンテナー レジストリで直ちに実行可能な高パフォーマンス ソフトウェアを提供し、さらに、ディープラーニング フレームワークの機能を強化することでオープンソース コミュニティに貢献しています。

NVIDIA では、フレームワーク、ドライバー、ハードウェアの新しいバージョン リリースに合わせて継続的な改善と更新を行っています。ユーザーのテストや統合の負担を解消し、すべての要素が最適化されて最大のパフォーマンスを発揮できるように努めています。NGC コンテナー レジストリで提供されるディープラーニング ソフトウェアにより、データ サイエンティストや研究者は、さまざまな分野や業界で飛躍的な成果を挙げ、常に新しい重要課題を AI で解決できるようになります。

NGC の詳細については、次のサイトを参照してください。
www.NVIDIA.com/ja-jp/gpu-cloud

NGC には、次のサイトから無償で登録することができます。
www.nvidia.com/ngcsignup (英語)