# NVIDIA TensorRT 3、ハイパースケール データ センターでの AI 推論を劇的に加速

*Alibaba*、*Baidu*、*Tencent*、*JD.com*、*Hikvision* が
プログラマブルな推論の高速化に向け *NVIDIA TensorRT* を採用

**中国・北京 — GTC China — (2017 年 9 月 26 日) —** NVIDIA (NASDAQ: NVDA) はこのたび、新しい NVIDIA® TensorRT 3 AI 推論ソフトウェア (英語) を発表しました。本ソフトウェアは、自動運転車やロボットをはじめ、クラウドからエッジ デバイスまで、推論のパフォーマンスの大幅な向上とコスト削減を可能にします。

TensorRT 3 を NVIDIA GPU と組み合わせることで、画像認識や音声認識、自然言語処理、画像検索、個人に合わせたレコメンデーションなど、AI 対応サービスのあらゆるフレームワークで超高速の効率的な推論を実行できるようになります。TensorRT と NVIDIA Tesla® GPU アクセラレータは、CPU よりも最大 40 倍の高速化[(1)] を、CPU ベースのソリューションの 10 分の 1 のコスト[(2)] で達成します。

NVIDIA の創設者兼 CEO であるジェンスン・フアン (Jensen Huang) は、次のように述べています。「多くのインターネット企業が先を争って何十億人ものユーザーを抱えるサービスへの AI 導入を進めていることから、AI 推論のワークロードが急増しています。NVIDIA TensorRT は、世界初のプログラマブルな推論アクセラレータです。CUDA のプログラマビリティによって、TensorRT は多様性と複雑さが増すディープ ニューラルネットワークを加速できます。また、TensorRT によって劇的な高速化を達成することで、サービス プロバイダーは、このような演算負荷の高い AI ワークロードを低コストで展開できるようになります。」

さまざまな業界で 1,200 社を超える企業が、データから新たな洞察を引き出してインテリジェント サービスを企業や消費者に展開するため、NVIDIA の推論プラットフォームをすでに利用し始めています。それらの企業には、Amazon、Microsoft、Facebook、Google のほか、Alibaba、Baidu、JD.com、iFLYTEK、Hikvision、Tencent、WeChat といった中国の大手企業も含まれています。

SAP の最高情報責任者であるユルゲン・ミュラー (Juergen Mueller) 氏は、次のように述べています。「TensorRT ソフトウェアを Tesla GPU 上で使用する NVIDIA の AI プラットフォームは、SAP の高まる推論への要件に対する取り組みの中心となる、卓越したテクノロジです。TensorRT と NVIDIA GPU によって、お客様のニーズを満たすために機械学習のパフォーマンスと汎用性を最大化させ、サービスをリアルタイムで提供することが可能になります。」

JD の AI およびビッグ データ担当シニア ディレクターであるアンディ・チェン (Andy Chen) 氏は、次のように述べています。「JD.com は、自社のデータ センターでの推論に NVIDIA の GPU とソフトウェアを利用しています。NVIDIA の TensorRT を Tesla GPU 上で使用することで、サーバーの数を 20 分の 1 に抑えながらも、1,000 件の HD ビデオ ス

トリームの推論をリアルタイムで同時に実行できるようになりました。NVIDIA のディープラーニング プラットフォームは、卓越したパフォーマンスと効率性を JD にもたらしています。」

TensorRT 3 は、AI アプリケーションの運用展開を可能にする、ハイパフォーマンスの最適化コンパイラおよびランタイム エンジンです。ハイパースケール データ センター、組み込みまたは車載用 GPU プラットフォームに対する、推論用にトレーニングされたニューラルネットワークの最適化、検証、展開を迅速に行うことができます。

非常に高精度の INT8 および FP16 でのネットワークの実行が可能になるため、データ センターの運営担当者は、調達コストや年間のエネルギー コストを数千万ドル単位で節約できます。また、開発者が利用すれば、トレーニング済みのニューラルネットワークを取り入れて、たった 1 日で既存のトレーニング フレームワークよりも 3 ～ 5 倍高速の展開可能な推論ソリューションを開発できるようになります。

NVIDIA は、AI のさらなる高速化に向けて次のようなソフトウェアを新たに導入しました。

- **DeepStream SDK:** NVIDIA DeepStream SDK (英語) によって、リアルタイムの低遅延ビデオ分析を大規模に実行できます。開発者は、INT8 の精度や GPU アクセラレーテッドのコード変換といった高度なビデオ推論機能を組み込み、オブジェクトの分類や状況の理解などの AI を利用したサービスをサポートして、単一の Tesla P4 GPU アクセラレータ (英語) で最大 30 件の HD ストリームをリアルタイムで処理できるようになります。

- **CUDA 9:** NVIDIA のアクセラレーテッド コンピューティング ソフトウェア プラットフォームである CUDA® (英語) の最新バージョン。NVIDIA Volta アーキテクチャベースの GPU、最大 5 倍高速なライブラリ、スレッド管理用の新しいプログラミング モデル、デバッグ ツールやプロファイリング ツールの更新をサポートして、HPC アプリケーションやディープラーニング アプリケーションを高速化します。CUDA 9 は、Tesla V100 GPU アクセラレータで最大のパフォーマンスを発揮できるよう最適化されています。

**データ センターでの推論**

データ センターの責任者は、すべてのサーバーで最大限の生産性を維持するため、絶えずパフォーマンスと効率のバランスを取っています。Tesla GPU アクセラレーテッド サーバーなら、ディープラーニングによる推論を利用するアプリケーションやサービスで 100 台のハイパースケール CPU を置き換え、貴重なラック スペースを解放して、エネルギーや冷却の要件を低減し、90% ものコストを削減できます。

NVIDIA Tesla GPU アクセラレータは、ディープラーニング推論ワークロードでの最大のスループット、最高の効率性、最小限の遅延を兼ね備え、AI による新たな体験を実現できる、最適な推論ソリューションとなります。

**自動運転車や組み込みアプリケーションでの推論**

NVIDIA の統合アーキテクチャによって、あらゆるディープラーニング フレームワークのディープ ニューラルネットワークのトレーニングをデータ センター内の NVIDIA DGX™ Systems で実行し、ロボットから自律走行車まで、あらゆる種類のデバイスに展開できるため、エッジでの推論をリアルタイムで行えるようになります。

自律型トラック輸送テクノロジ (英語) の開発を手がけるスタートアップ企業である TuSimple は、TensorRT による最適化を利用して推論パフォーマンスを 30% 向上させました。6 月には、NVIDIA GPU とカメラを主なセンサーとして使用し、カリフォルニア州サンディエゴからアリゾナ州ユマまでの 170 マイル (約 270 km) のレベル 4 テスト走行を成功させています。TensorRT によってパフォーマンスを向上させることで、TuSimple は、より多くのカメラ データを分析して、応答時間を損なわずに自社の自律型トラックに新しい AI アルゴリズムを取り入れることができました。

**NVIDIA についての最新情報:**
公式ブログ NVIDIA blog、Facebook、Google+、Twitter、LinkedIn、Instagram、NVIDIA に関する動画 YouTube、画像 Flickr

**NVIDIA について**
NVIDIA が 1999 年に開発した GPU は、PC ゲーム市場の成長に拍車をかけ、現代のコンピューターグラフィックスを再定義し、並列コンピューティングを一変させました。最近では、GPU ディープラーニングが最新の AI、つまりコンピューティングの新時代の火付け役となり、世界を認知して理解できるコンピューター、ロボット、自動運転車の脳の役割を GPU が果たすまでになりました。今日、NVIDIA は「AI コンピューティングカンパニー」として知名度を上げています。詳しい情報は、http://www.nvidia.co.jp/ をご覧ください。

(1) TensorRT 3 RC を実行する NVIDIA Tesla V100 GPU での ResNet-50 のパフォーマンスと、Intel Xeon-D 1587 Broadwell-E CPU および Intel DL SDK での ResNet-50 のパフォーマンスとの比較による。AVX512 を搭載した Skylake で 2 倍のパフォーマンスを達成したという Intel の公式な主張を踏まえて、スコアを 2 倍にした。
(2) 8 基の NVIDIA Tesla V100 を搭載した HGX-1 サーバーのコストおよび ResNet-50 の推論パフォーマンスと、デュアル ソケットの Intel Skylake スケールアウト サーバーのコストおよびResNet-50 のパフォーマンスの推定値との比較による。Skylake のパフォーマンスの推定値は、AVX512 を搭載した Skylake で 2 倍のパフォーマンスを達成したという Intel の公式な主張を踏まえた。