NVIDIA、ディープラーニング推論の能力を
ハイパースケール データセンターに拡大

NVIDIA TensorRT 4、TensorFlow への統合、Kaldi 音声認識のアクセラレーション、
ONNX サポートの拡大を発表、GPU 推論が最大で CPU の 190 倍高速に

米国カリフォルニア州サンノゼ — GPU テクノロジ カンファレンス — (2018 年 3 月 27 日) — NVIDIA (NASDAQ: NVDA) は本日、潜在的な推論の市場を世界で 3,000 万台のハイパースケール サーバーに拡大すると同時に、ディープラーニングを活用したサービスの提供コストを大幅に低減する、一連の新しいテクノロジとパートナーシップを発表しました。

NVIDIA の創業者兼 CEO であるジェンスン・フアン (Jensen Huang) は、GTC 2018 の基調講演において、ディープラーニング推論向けの GPU アクセラレーションがどれほど勢いを増しているかを説明しました。GPU アクセラレーションは、音声認識、自然言語処理、レコメンダー システム、画像認識などの能力を新たにサポートし、データセンターや自動車用アプリケーション、ロボットやドローンなどの組み込みデバイスでも利用されています。

NVIDIA は、新しいバージョンの TensorRT 推論ソフトウェアと、広く使われている Google の TensorFlow フレームワークへの TensorRT の統合を発表しました。また、最も広く使われている NVIDIA の音声認識フレームワークである Kaldi が GPU 向けに最適化されたことも発表しました。Amazon、Facebook、Microsoft などのパートナーと NVIDIA の緊密なコラボレーションにより、ONNX および WinML を利用した GPU アクセラレーションを開発者がより容易に利用できるようになります。

NVIDIA のバイス プレジデント兼アクセラレーテッド コンピューティング担当ゼネラル マネージャーであるイアン・バック (Ian Buck) は、次のように述べています。「実働するディープラーニング推論向けに GPU アクセラレーションを利用すると、大規模なニューラルネットワークでもリアルタイムかつ最低のコストで稼働させることができます。私たちは、より多くのインテリジェントなアプリケーションやフレームワークに対するサポートを迅速に拡大させたことにより、ディープラーニングの品質を向上させ、3,000 万台のハイパースケール サーバーのコスト削減に貢献できるようになりました。」

TensorRT、TensorFlow への統合

NVIDIA は、ディープラーニング推論を幅広いアプリケーションにおいて加速する TensorRT 4 ソフトウェアを発表しました。TensorRT により、INT8 および FP16 のネットワーク実行をきわめて高い精度で提供できるようにするだけでなく、データセンターのコストを最大 70% 削減できます。(1)

TensorRT 4 は、ハイパースケール データセンター、組み込みおよび自動車用の GPU プラットフォームにおいてトレーニングされたニューラルネットワークを迅速に最適化、検証、展開するために利用できます。このソフトウェアは、コンピューター ビジョン、ニューラル機械翻訳、自動音声認識、音声合成、レコメンデーション システムなど、一般的なアプリケーションに対し、CPU と比較して最大 190 倍高速なディープラーニング推論を実現します。(2)

さらに開発を合理化するため、NVIDIA と Google のエンジニアは、TensorRT を TensorFlow 1.7 に統合し、ディープラーニングの推論アプリケーションを GPU で実行しやすくしました。

Google のエンジニアリング ディレクターであるラジャット・モンガ (Rajat Monga) 氏は、次のように述べています。「TensorFlow のチームは、NVIDIA ととても緊密に協力しており、NVIDIA GPU において可能な限り最高のパフォーマンスをディープラーニング コミュニティーにもたらしています。TensorFlow の NVIDIA TensorRT との統合により、Volta Tensor Core テクノロジを搭載した NVIDIA ディープラーニング プラットフォームにおいて、(低いレイテンシ目標内で GPU を通常実行する場合と比較して) 最大で 8 倍高速な推論のスループットを提供できるようになり、TensorFlow 内で GPU 推論のパフォーマンスを最大にすることが可能になりました。」

NVIDIA は、世界最先端の音声フレームワークである Kaldi を最適化し、GPU において、より高速なパフォーマンスを実現してきました。音声認識の GPU アクセラレーションは、消費者には、より精度の高い、有用なバーチャル アシスタンスを実現し、データセンターのオペレーターには、より低額なデプロイメント コストを実現します。

広範な業界をサポート

世界中のさまざまな企業の開発者は、データから新しい洞察を発見したり、インテリジェントなサービスを企業や顧客に提供したりするために、TensorRT を使用しています。

NVIDIA のエンジニアが Amazon、Facebook、および Microsoft と密接に連携して作業した結果、Caffe 2、Chainer、CNTK、MXNet、および Pytorch などの ONNX フレームワークを使用する開発者は、NVIDIA ディープラーニング プラットフォームに容易にデプロイすることが可能になりました。

SAP の機械学習部門を率いるマーカス・ノガ (Markus Noga) 氏は、次のように述べています。「ディープラーニングに基づくレコメンド アプリケーションを NVIDIA Tesla V100 GPU で実行する TensorRT を評価したところ、CPU ベースのプラットフォームと比べて推論の速度とスループットが 45 倍向上しました」

Twitter Cortex の代表であるニコラス・カムチャツキー (Nicolas Koumchatzky) 氏は、次のように述べています。「GPU を使用することにより、プラットフォームでメディアを理解することが可能になりました。メディアのディープラーニング モデルのトレーニング時間が大幅に短縮するだけでなく、推論時にライブ ビデオのリアルタイムの理解を導出することができるからです」

また、Microsoft は最近、Windows 10 アプリケーションの AI サポートを発表しましたが、NVIDIA は Microsoft と提携して GPU アクセラレーション ツールを構築しました。このツールは、開発者がよりインテリジェントな機能を Windows アプリケーションに組み込むことができるよう支援するものです。

さらに、NVIDIA では、Kubernetes 向けの GPU アクセラレーションを発表しました。これにより、企業はマルチクラウド GPU クラスターで推論をデプロイするのが容易になります。NVIDIA は、Kubernetes のエコシステムをサポートするため、オープンソースのコミュニティーに対する GPU の拡張に貢献しています。

それに加え、MATLAB ソフトウェアのメーカーである MathWorks は、TensorRT と MATLAB の統合を発表しました。これでエンジニアや科学者は、NVIDIA DRIVE™、Jetson™、および Tesla® プラットフォーム用に MATLAB が提供する高性能の推論エンジンを自動生成できます。

データセンターでの推論

データセンターのマネージャーは、サーバー群の生産性を最大限に保つため性能と効率のバランスを常に保ちます。NVIDIA Tesla GPU アクセラレーション サーバーは、ディープラーニングの推論アプリケーションおよびサービス用の CPU サーバーのラックの代わりとなり、貴重なラック スペースを解放し、エネルギーと冷却の要件を減らすことができます。

自動運転車、埋め込みプラットフォームでの推論

また、TensorRT は NVIDIA DRIVE 自律走行車や NVIDIA Jetson 埋め込みプラットフォームにもデプロイできます。あらゆるフレームワーク上のディープ ニューラルネットワークは、データセンター内の NVIDIA DGX™ システムでトレーニングされた後、ロボットから自律走行車に至るまですべてのタイプのデバイスにデプロイされ、末端でリアルタイムの推論を実行できます。

TensorRT を使用すると、開発者は推論のデプロイのための性能調整よりも、新しいディープラーニングを利用したアプリケーションの開発に集中できます。開発者は TensorRT を利用して、INT8 または

FP16 の精度を使用した超高速の推論を実現でき、これによりレイテンシが大幅に短縮されます。これは、組み込みプラットフォームや自動車プラットフォームでのオブジェクトの検出やパス プランニングなどの機能に不可欠です。

NVIDIA Developer Program のメンバーは、https://developer.nvidia.com/tensorrt で TensorRT 4 リリース候補版の詳細を参照できます。

(1) 主要なクラウド サービス プロバイダーの代表的な混合ワークロードに基づく総所有コスト: Neural Collaborative Filtering (NCF) が 60 パーセント、Neural Machine Translation (NMT) が 20 パーセント、Automatic Speech Recognition (ASR) が 15 パーセント、コンピューター ビジョン (CV) が 5 パーセント。ソケットあたりのワークロードの高速化 (Tesla V100 GPU と CPU を比較): NCF が 10 倍、NMT が 20 倍、ASR が 15 倍、CV が 40 倍。CPU ノード構成は 2 ソケットの Intel Skylake 6130。GPU の推奨ノード構成は 8 基の Volta HGX-1。

(2) 性能の向上はさまざまな重要なワークロードで観測されています。たとえば、レイテンシが 7 ミリ秒の ResNet50 v1 の推論性能は、レイテンシが最小 (バッチ=1) の単一ソケットの Intel Skylake 6140 で TensorFlow を使用する場合よりも、Tesla V100 GPU で TensorRT を使用した方が 190 倍高速になります。

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。

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

本件に関するお問い合わせ先:

エヌビディア　広報/マーケティングコミュニケーションズ
中村かおり　Email アドレス：knakamura@nvidia.com　TEL: 03-6743-8712

吉川香葉子　Email アドレス：kyoshikawa@nvidia.com TEL: 080-8891-3352

このプレスリリースに含まれる特定の記述は将来の見通しに関する記述であり、リスクや不確実性の影響を受けるため、結果が予測とは著しく異なる可能性があります。その記述には次のようなものが含まれますが、これに限定されません。NVIDIA TensorRT 4 およびそれを TensorFlow フレームワークに統合することの利点、影響、性能、使用、および能力。利用が加速しているディープラーニングの推論用の GPU アクセラレーションおよびその影響と利点。NVIDIA が Kaldi と連携して GPU を最適化すること、NVIDIA がパートナーと提携すること、TensorRT が世界中で使用されること、Kubernetes 向けの GPU アクセラレーション、NVIDIA Tesla GPU アクセラレーション サーバー、および MATLAB ソフトウェアと TensorRT の統合の利点と影響。NVIDIA のテクノロジが潜在的な推論市場を拡張していること。ディープラーニングの質を向上させ、ハイパースケール サーバーのコストを削減するインテリジェントなアプリケーションおよびフレームワークのサポート。かかるリスクと不確実性は、世界的な経済環境、サードパーティに依存する製品の製造・組立・梱包・試験、技術開発および競合による影響、新しい製品やテクノロジの開発あるいは既存の製品やテクノロジの改良、当社製品やパートナー企業の製品の市場への浸透、デザイン・製造あるいはソフトウェアの欠陥、ユーザーの嗜好および需要の変化、業界標準やインターフェイスの変更、システム統合時に当社製品および技術の予期せぬパフォーマンスにより生じる損失などを含み、その他のリスクの詳細に関しては、2018 年 1 月 28 日を末日とする会計期間の Form 10-K レポートなど、米証券取引委員会 (SEC) に提出されている NVIDIA の報告書に適宜記載されます。SEC への提出書類は写しが NVIDIA のウェブサイトに掲載されており、NVIDIA から無償で入手することができます。これらの将来予測的な記述は発表日時点の見解に基づくものであって将来的な業績を保証するものではなく、法律による定めがある場合を除き、今後発生する事態や環境の変化に応じてこれらの記述を更新する義務を NVIDIA は一切負いません。