6. 医療分野におけるAI開発のためのGPU技術の現状と展望

鈴木 博文 エヌビディア合同会社エンタープライズ事業部

2018年3月末に米国・サンノゼで開催したNVIDIA主催のカンファレンス（GPU Technology Conference：GTC）での最新情報をベースに，最新のVoltaアーキテクチャを採用したGPUアクセラレータの「Tesla」，このTeslaを搭載したGPUワークステーションの「DGX Station」，GPUサーバーの「DGX-1」，新規開発の「DGX-2」といったハードウエアに加え，NVIDIA GPU Cloud（NGC）をはじめとするソフトウエアの最新動向も紹介する。

最後に，今後の展望として，NVIDIAが開始したメディカルイメージングのためのプロジェクト“Clara”についても紹介する。

NVIDIAのAI研究・開発環境強化の取り組み

自動運転やエンタープライズ，ロボティクスなどの領域で人工知能（AI）の研究および開発は加速する一方であり，画像系で多用される畳み込みニューラルネットワーク（convolutional neural network：CNN）に加え，再帰型ニューラルネットワークネットワーク（recurrent neural network：RNN），敵対的生成ネットワーク（generative adversarial network：GAN），強化学習，さらに新しいネットワークも加わり，AIモデルはますます多様化，大規模化が進んでいる（**図1**）。

NVIDIAのコア技術であるGPUは，グラフィックス処理の高速化という成り立ちから，並行処理やディープラーニングで多用される積和演算に強く，今日のAIの研究・開発では必須となっており，ここ数年のAI研究の加速化に伴い，GPUの活用範囲および処理性能への要求は拡大する一方となっている。

NVIDIAでは，この要求を受けてGPUの性能強化を継続しており，2013年時点のFermi GPUを搭載したサーバーから2018年のVolta GPU搭載サーバーへの推移に見られるように，ムーアの法則を超えるペースでの性能強化を続けている（**図2**）。

NVIDIAの代表的なGPU技術について具体的に紹介したい。NVIDIAのGPUアクセラレータの「Tesla V100」は，最新のGPUアーキテクチャVoltaを採用し，ディープラーニング用に新たに開発された640個のTensor Coreを備え，演算能力は倍精度7.8 TeraFLOPS，単精度15.7 TeraFLOPS，ディープラーニング性能は125 TeraFLOPSと，初めて100 TeraFLOPSを超えた。NVLinkによるGPU間接続も300GB/秒と前世代の「Tesla P100」（Pascalアーキテクチャ）の2倍，PCIe Gen3×16相互接続との比較では10倍の速度となった。この結果，ディープラーニングの学習演算能力は，1年前に最高性能だったTesla P100よりも3倍程度強化され，学習，推論共に大幅な時間短縮が可能となった。GPUメモリも32GBと倍増され，これまでよりも大きなモデル，データを扱うことが可能となった。

このTesla V100を8基搭載したラックタイプサーバーのDGX-1では，ディープラーニング性能は，ついに1 Peta FLOPSと“Peta”レベルを達成した。このVoltaアーキテクチャのDGX-1は，前世代のTesla P100で15時間を要したResNet-50，90エポックの学習が，わずか5時間ほどで終了するなど，ディー

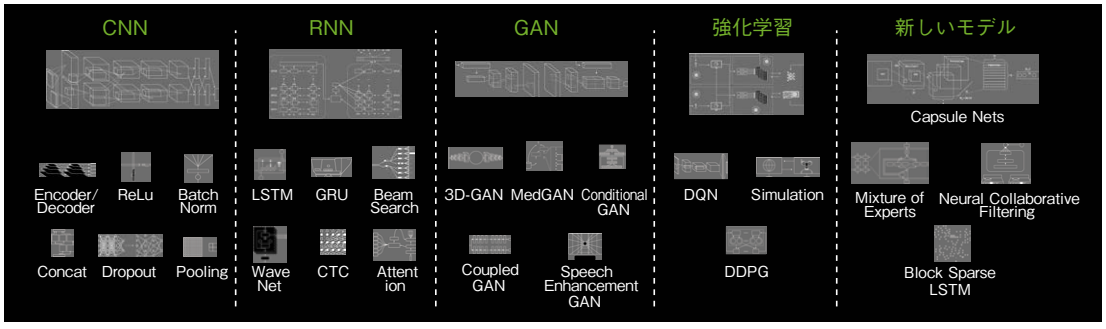

図1 多様化するAIモデル

〈0913-8919/18/￥300/ 論文 /JCOPY〉