特集 AIで広がる医療の可能性

Ⅱ　生成AIの医療応用への展望

3. エヌビディアにおける生成AIの取り組みと医療分野での展開

山田　泰永　エヌビディア（同）ヘルスケア開発者支援シニアマネージャー

本稿では，まず大規模言語AIモデルや生成AI全般に向けたエヌビディアのさまざまな取り組みやソリューションを紹介し，その上でヘルスケア・医療分野に特化した取り組みや展望を述べたい。

ハードウエア面での取り組み

まず，GPUのハードウエアについて，GPUは元来グラフィック計算処理のために発展してきたプロセッサであり，グラフィック処理には32ビット浮動小数点計算が用いられてきた。AI計算においても当初は32ビット浮動小数点計算が用いられていたが，その後の発展過程において16ビット浮動小数点や8ビット浮動小数点，また，特に画像領域では8ビット整数計算でも精度に大きなインパクトを与えずに高速な処理が可能となってきた。当社においても，まずGPUでこれらの16ビットや8ビット浮動小数点計算をサポートすると同時に，AI計算の大半を占める行列の積和演算において，4×4の行列を一度に計算できるTensorCoreという機構を実現し，実装してきた。さらに現在，最新の大規模言語モデルを支えるTransformer型AIモデルでは，4ビットまでビット数を落としても精度をほぼ維持できる部分も存在するということで，最新のH100 GPUでは4ビットでの高速計算をサポートし，また，4ビット計算化が可能な部分を検出して自動で適用するソフトウエアと組み合わせた「Transformer Engine」という機能も搭載している。

大規模生成AIモデルのサイズはとどまることを知らずに拡大を続けており，これを収容し高速処理するために，当社では率先してHBMを中心とする高速で大容量なメモリを採用してきている。さらに，最近の超大規模モデル開発では，マルチGPU，マルチノードでの分散並列処理がほぼ必須となっており，当社の「NVLink」や「NVSwitch」による超高速なGPU間通信，ノード間通信は，これをハードウエア面から下支えしている。

ソフトウエア面での取り組み

ソフトウエア側については，ディープラーニングAI以前から，当社では長年にわたってGPUで高速に汎用計算を実行する基盤である「CUDA」を提供し発展させてきたが，ディープラーニングAIの普及期からは，その計算をGPU/CUDA環境上で高速に実行するために，CuDNNをはじめとするライブラリ群を無償で提供するとともに，AIモデル開発の事実上の標準言語環境たるPyTorchなどのフレームワークの開発者と緊密に連携して各種ソフトウエアの最適化を行ってきた。そのため，GPUに向けた特殊なコードやスクリプトの改変をすることなく，多様な最新のAI計算を透過的に高速処理できるようになった。そして，それを進化させ続けてきたことが，当社のGPUと関連ソフトウエアが，現在先進的なAIコンピューティングにおける事実上の標準に近い環境として普及した背景である。

改めて，大規模言語モデルや生成AIの領域では，大量のパラメータを含む大規模なモデルで大量のデータを学習する必要があるため，マルチGPUおよびマルチノードで大規模にモデル並列，データ並列を行い，分散並列学習することがほぼ必須になっている。こうしたニーズに対応して当社では，PyTorchフレームワーク上で大規模なTransformer型モデルを実現するための基本的な機能ブロックと，大規模分散並列学習を容易にするライブラリとしてMegatron-Coreを提供している。さらに，Megatron-Coreをベースに，BERT，GPT，T5といったさまざまな実際の大規模モデルを容易に分散並列学習するためのMegatron-LMも提供している。Megatron-LMは，そのままで開発者に活用されるとともに，さまざまなサードパーティツール群に利用されたり影響を与えたりしており，特に現在，開発者の間で幅広く活用されているマイクロソフト社のMegatron-DeepSpeedにおいても活用され，言わば「縁の下の力持ち」として大規模生成AIの研究開発を支えていると言えるだろう。

これらの基盤ソフトウエアツール群だけにとどまらず，当社では，最新の大規模言語モデル・生成AIを容易に評価，データキュレーションを経たファインチューニングやカスタム開発，そして配備を行うための統合環境として，「NeMo」フレームワークを公開している。さらに，

〈0913-8919/24/￥300/論文/JCOPY〉