IoT ビッグデータ処理用 AI 計算機構築のためのデバイス技術

Device Technologies to Construct AI for IoT Big-Data Process

産業技術総合研究所　○大内　真一

National Institute of AIST ○Shin-ichi O'uchi

E-mail: shinichi.ouchi@aist.go.jp

【はじめに】IoT によって生成された大量のデータをクラウドで処理・学習する人工知能の必要性が増している。これに対して、これまで計算機の高性能化を支えてきた CMOS スケーリングは一層困難になってきており、計算機アーキテクチャ・プログラミング技術・素子技術の広い階層にわたる統合的技術開発によらなければ、計算の規模・速度の拡大はもはや不可能と考えられる時代となった。本講演では、一例として深層学習[1]に関して考察し、計算機アーキテクチャとデバイス技術のコラボレーションの可能性について議論する。

【計算回路の専用化によるエネルギー効率向上】Google の Tensor Processing Units[2]や MIT-NVIDIA の Eyeriss[3]など、計算回路の専用性を高める動きがクラウド・エッジの両方で起きている。これらは、近年の CMOS スケーリングの困難性と無関係ではない。すなわち、専用計算回路をチップ内部に実装することは、2000 年代初頭よりの電圧スケーリング停滞に端を発するダークシリコン問題[4]をエネルギー効率の高い専用回路によって有効活用することに通じる。人工知能の演算に対する需要が増していけば、専用チップの設計コストを吸収しつつエネルギー効率の問題を解決し、演算の大規模化に成功する可能性がある。さらに深層学習がデータ駆動型の処理であることから、プログラミングコスト低減に通じる。

【深層学習における演算性能とメモリバンド幅、通信速度】深層学習に特化した計算システムを構築するためにディープニューラルネット(DNN)の計算アルゴリズムを考えると、前半の畳み込み層では積和演算数が大量に必要であり、演算要素を多数集積し並列度を高めることが有効である。反対に後半の全結合層では、積を得るための定数のメモリロードが大量に必要となり、メモリバンド幅が計算速度を律速する。さらに計算規模を大規模化するためには、チップ間をまたがる複数要素間のデータ転送ボトルネックを解消する必要がある。

【深層学習を加速するデバイス技術】上記に照らし今後開発を加速すべきデバイス技術要素は、(1)TSV、磁界結合 3 次元集積、さらにはモノリシック三次元集積: 単一キャッシュの周辺で積和算を大スループットで行うモジュールもしくは HBM(High Bandwidth Memory)の実現、(2) ストレージクラスメモリのオンロジック集積: HBM の低エネルギー・大容量化、(3) 集積フォトニクス: モジュール間データ転送の高速化、などが挙げられる[5]。需要の高い計算問題特徴的なボトルネックを解消するデバイス技術の開発は、IoT を支える AI に大きく寄与すると期待される。

【謝辞】本講演の一部は、神戸大学川口博教授、東京大学工藤知宏教授、産業技術総合研究所高野了成グループ長、松川貴グループ長並びに入沢寿史氏、更田裕司氏との議論に基づきます。感謝申し上げます。

【参考文献】

[1] A. Krizhevsky *et al.*, Proc. Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012.

[2] https://cloudplatform.googleblog.com/2016/05/Google-supercharges-machine-learning-tasks-with-custom-chip.html

[3] Y.H. Chen *et al.*, 2016 IEEE International Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers, pp. 262-263, 2016.

[4] M.B. Tayler, IEEE Micro 33, No. 5, pp. 8-19, 2014.

[5] https://unit.aist.go.jp/rai/star/impulse/impulse.html

© 2016年 応用物理学会