

AI を加速する不揮発性メモリを用いたアナログ積和演算への期待

Neural Network Accelerator Using Non-Volatile Memory as Synapse

for Next-Generation AI Hardware Platform

日本アイ・ビー・エム 東京基礎研究所, 石井 正俊

IBM Research - Tokyo, Masatoshi Ishii

E-mail: ishiim@jp.ibm.com

深層学習をはじめとする人工ニューラルネットワーク (ANN) を用いた人工知能 (AI) 技術はこの数年で目覚ましい発展を遂げている。今後、より高度な AI 機能を実現させるため、さらに大量のデータを用い、より多層で複雑なネットワーク構成を用いた計算処理が必要とされることが予想され、AI を支えるハードウェアプラットフォームもそのような膨大な計算量を効率よく処理することが強く求められる。現在の ANN における計算処理は、多くの部分を積和演算が占め、積和演算の高速化・低消費電力化が ANN の演算効率の向上に直結する。大量の浮動小数点演算器を有する GPU はこれら積和演算を同時並列に処理することを得意とし、機械学習の発展に大きく貢献してきた。近年では、AI に特化したプロセッサを開発して、アルゴリズムの改善に加え、重み表現ビット数削減、データフロー最適化など様々な手法を用い、演算効率の向上や低消費電力化をさらに加速させようとする試みも盛んに行われている[1]。

一方、指数関数的に増大すると予想される ANN の計算量を今後も持続的に支えていくためのハードウェアプラットフォームとして、新しいコンピューティングアーキテクチャにより飛躍的に演算効率を改善しようとする研究も活発に議論されている。その一つとして注目される技術が、resistive RAM (ReRAM)や phase change memory (PCM) などの不揮発性メモリを利用したアナログ積和演算である[2]。これらの不揮発性メモリは、もともとバイナリの (もしくは数レベルの) 値を格納するために開発されたデジタルメモリであるが、物性として本来持つアナログのコンダクタンス値を ANN の重みとしてメモリ素子に書き込むことにより、1 素子で 1 つの重みを表現することが可能である。また、アナログ集積回路技術を用いて、素子を通過した電流を加算することで、積和演算を短期間で同時並列に行う事ができる。このとき、フォン・ノイマン型コンピュータで必要とされるメモリ・プロセッサ間のデータ転送を行う必要はなく、全てのアナログ積和演算をメモリの直近で実行することができるため、フォン・ノイマン・ボトルネックを解消することができる。

本講演では、不揮発性メモリを用いたアナログ積和演算について、その利点やデバイスに要求される特性、スパイキングニューラルネットワークへの応用について紹介する。

参考文献

- [1] Norman P. Jouppi, et al., “In-Datacenter Performance Analysis of a Tensor Processing Unit”, ISCA '17.
- [2] G. Burr, et al., “Neuromorphic computing using non-volatile memory”, Advances in Physics: X 2:1, 89-124, 2017.