

不揮発性メモリデバイスを用いたニューラルネットワーク集積回路

Neural Network VLSI using Non-Volatile Memory Devices

日本アイ・ビー・エム株式会社 東京基礎研究所 岡崎篤也

IBM Research - Tokyo, Atsuya Okazaki

E-mail: a2ya@jp.ibm.com

データセンタに代表されるようなクラウドコンピューティングか、スマートフォンやセンサなどのエッジコンピューティングかは問わず、画像認識、自然言語処理、音声認識など多くのアプリケーションで、いわゆる人工知能と呼ばれる機能や信号処理のための計算の需要が近年激増している。その多くはニューラルネットワークを基とした様々なアルゴリズムが開発、利用展開されている。ニューラルネットワークは、ニューロンからニューロンへの入力信号と、そのニューロン間を接続するシナプス結合の重みとの積和演算が必要な計算負荷の内の大部分を占める。ネットワークの性能はニューロン数を増やすと高度になる傾向があり、高性能なニューラルネットワークを用いたサービスや機能を実装するためには、クラウドコンピューティングにおいてもエッジコンピューティングにおいても、大量の積和演算を高速かつ低消費電力で行うことが重要となる。

現在、このニューラルネットワークで必要となる積和演算のために、クラウドコンピューティングでは積和演算器を多数並列に搭載した Graphic Processing Unit (GPU) や Application Specific Integration Circuit (ASIC) が利用されており、エッジコンピューティングでも積和演算機を並列に内蔵したセンサや ASIC が利用され始めている。これらの積和演算器は Complementary Metal–Oxide–Semiconductor (CMOS) デジタル回路技術が利用されており、アルゴリズムの工夫により必要となる演算ビット精度を少なくすることができるようになってきているものの、それでも CMOS 回路で積和演算器を実装するには、ビット精度に対し数桁多い数のトランジスタが必要となる。またその演算器に入力されるニューロン間の信号やシナプスの重み、演算結果を保持するための Static Random-Access Memory (SRAM) や Dynamic Random-Access Memory (DRAM) などのデジタルメモリデバイスも大量に必要な上に、演算器とメモリの間のデータの移動が演算のパフォーマンスボトルネックになったり、システム全体の消費電力の主要因の一つになっている。

従来のデジタルメモリでは同時には一つか二つ程度のメモリデバイスしかアクセスされないという特性を活かし、2次元アレイ状に高密度に集積することが可能で、半導体製造プロセステクノロジーの進化に沿ってシリコン面積当たりの集積ビット数を伸ばしてきた。デジタルメモリは電圧または電荷が閾値を超えるかで1と0を区別しビットとして記憶保持し、複数ビットをまとめ一つの数値を表現する。このメモリデバイスを連続的な値を保持可能な新しい不揮発性アナログメモリに置き換えることにより、シナプスの重みを表す一つの値を単一メモリセルで保持することが可能になり、重みの記憶密度を劇的に向上できる可能性がある。また、従来の積和演算機ではニューロン間の信号や重みのデータや演算結果をメモリに逐次的に入出力していたのに対し、連続的な重みを高密度に保持している2次元メモリアレイ自体にニューロン間の信号を並列に入力することにより、その入力信号と重みの大規模並列なアナログ積和演算が可能になる。この様にデータを移動せずにメモリアレイ上で演算する in-memory computing により、従来の計算機アーキテクチャでの演算器とメモリの間のパフォーマンスボトルネックを解消すると共に、データの移動に起因する消費電力を削減できる可能性がある。

本発表では、このような新しい不揮発性のアナログメモリデバイスを用いた、単位消費電力当たりのニューラルネットワークとしての性能を向上することを目的とした集積回路の研究事例を紹介する。