論 文 内 容 の 要 旨

博士論文題目

Efficient Stochastic Computing Architectures for Deep Neural Networks
（深層ニューラルネットワークのための効率的な確率的計算アーキテクチャ）

氏 名 Nguyen Van Tinh

This thesis comprises four parts, where the first one presents a novel architecture for univariate radial basis kernel computation employing stochastic computing. The univariate radial basic function is optimized using proposed simple stochastic logic circuits. I validated this approach by comparison with both Bernstein polynomial and two-dimensional finite-state machine-based implementation. Optimally, the mean absolute error is reduced by 40% and 80% compared to two other well-known approaches, Bernstein polynomial and two-dimensional finite-state machine-based implementation, respectively. In terms of hardware cost, our proposed solution required as much as the Bernstein method did. Moreover, the proposed approach outperforms the two-dimensional finite-state machine-based implementation, roughly 54% less hardware cost. Regarding the critical path delay, the proposed system is less than 12% than others on average. Besides, the proposed architecture also required 70% less power than two-dimensional finite-state machine-based implementation.

The second part of the thesis proposes a novel technique implementation of hyperbolic *tanh(ax)* and *sigmoid(2ax)* functions for high precision and compact computational hardware based on stochastic logic. This work demonstrates the stochastic computation of *tanh*(ax) and *sigmoid(2ax)* functions-based Bernstein polynomial using a bipolar format. The format conversion from bipolar to unipolar format is involved in our implementation. One achievement is that our proposed method is more accurate than the state-of-the-art, including the FSM-based method, JK-FF, and general unipolar division. On average, 90% of improvement of this work in terms of mean square error (MAE) has been achieved while the hardware cost and power consumption are comparable to the previous approaches.

The third and fourth parts of the thesis propose an in-memory binary spiking neural network (BSNN) using stochastic computing bitstreams for information encoding. The residual BSNN learning using a surrogate gradient that shortens the time steps in the BSNN while maintaining sufficient accuracy is proposed. At the circuit level, presynaptic spikes are fed to memory units through differential bit lines (BLs), while binarized weights are stored in a subarray of nonvolatile spin-transfer-torque magneto-resistive RAM (STT-MRAM). The hardware/software co-simulation results indicate that the proposed design can deliver a performance of 176.6 TOPS/W for an in-memory computing (IMC) subarray size of 1x288. The classification accuracy reaches 97.92% (83.85%) on the MNIST (CIFAR-10) dataset. The impacts of the device nonidealities and process variations are also thoroughly covered in the analysis.

氏　名	Nguyen Van Tinh

（論文審査結果の要旨）（Ａ４　１枚　１、２００字程度）

本論文は、非ノイマン型ハードウェアに関する研究である。スパイクエンコーディングとメモリスタ（STTRAM）を使用して、アナログとデジタルのハイブリッド演算を構成し、メモリ内バイナリスパイキングニューラルネットワークを提案している。第1に、ストカスティク・コンピューティングを用いたガウス関数計算のための新しいアーキテクチャを提案している。本手法は、Bernstein 多項式および 2 次元有限状態マシンベースの実装と比較して、平均絶対誤差をそれぞれ 40%および 80%削減することに成功した。第２に、ストカスティク・ロジックに基づく高精度かつ高効率ハードウェアのために，双曲線 tanh(ax)と sigmoid(2ax)関数の新しい実装手法を提案している。バイポーラ形式からユニポーラ形式への変換は、本実装に含まれる。評価の結果、FSM ベース法、JK-FF 法、一般的なユニポーラ分割法などの技術と比較して、平均二乗誤差(MAE)で 90%の改善を達成しつつ、ハードウェアコストと消費電力は従来のアプローチと同程度であることを示した。第３に、情報エンコーディングに確率計算ビットストリームを用いた、インメモリ型バイナリスパイキングニューラルネットワーク（BSNN）を提案している。また、十分な精度を保ちつつ BSNN の時間ステップを短縮する、代用勾配を用いた残余 BSNN 学習手法を提案している。回路レベルでは、シナプス前スパイクは差動ビット線（BL）を介してメモリユニットに供給され、2 値化された重みは不揮発性スピントランスファートルク磁気抵抗 RAM（STT-MRAM）のサブアレイに保存される。ハードウェア/ソフトウェア協調シミュレーションの結果、提案設計は、インメモリコンピューティング（IMC）サブアレイサイズ 1x288 で 176.6 TOPS/W の性能を発揮することが示された。また、MNIST と CIFAR-10 データセットに対して、それぞれの認識精度が 97.92%（83.85%)に達している。デバイスの非理想性やプロセスのばらつきの影響も徹底的に解析している。総じて、ハードウェアの小型化、高速化、低電力化を達成している。

以上、本論文は学術上、実際上寄与するところが少なくない。よって、本論文は博士（工学）の学位論文として価値あるものと認める。