次世代不揮発性メモリを用いたエッジ AI アクセラレータの開発に向けた取り組み

林越 正紀[†a)]

Initiatives to develop edge AI accelerator using next-generation non-volatile memory

Masaonri HAYASHIKOSHI[†a)]

あらまし　AI の演算には消費電力が大きいという課題があり，近年 AI アクセラレータの研究が盛んである．中でも，メモリ素子を用いたコンピューティング・インメモリ (Computing in Memory: CiM) 技術は，従来のデジタル AI アクセラレータより，10 ～ 100 倍の推論電力効率が得られるというメリットがある．本論文では，電力効率を最大限に活用できる CiM アーキテクチャの実現可能性について報告する．

キーワード　AI アクセラレータ，コンピューティング・インメモリ，CiM，エッジ AI，NVRAM

Abstract　AI computations have the problem of high power consumption, and research into AI accelerators has been active in recent years. Among these, computing in-memory (CiM) technology that uses memory devices has the advantage of being 10 to 100 times more efficient in inference power than traditional digital AI accelerators. In this paper, we discuss the feasibility of a CiM architecture that can maximize power efficiency.

Keywords　AI accelerator, Computing in-Memory, CiM, edge AI, NVRAM

1. まえがき

Society 5.0 においては，エッジ・インテリジェンスからはじまるイノベーションにより，クラウドやエッジでのビッグデータ活用だけでは解決できない領域で，より安全で健やかな暮らしを支える環境に優しいスマート社会の実現に貢献していくことが期待されている．

ここで，エッジでの AI とは，予め取得した学習データをエッジト側に供給し，負荷の小さい推論を実行する必要があるが，そのためには低消費電力化が重要となる．

AI の演算には，大量のデータ転送と演算が必要なため，消費電力が大きいという課題があり，近年 AI アクセラレータの研究が盛んである．中でも，メモリ素子を用いたコンピューティング・インメモリ (Computing in Memory: CiM) 技術は，従来のデジタル AI アクセラレータより，10 ～ 100 倍の推論電力効率が得られるというメリットがある（図 1）．

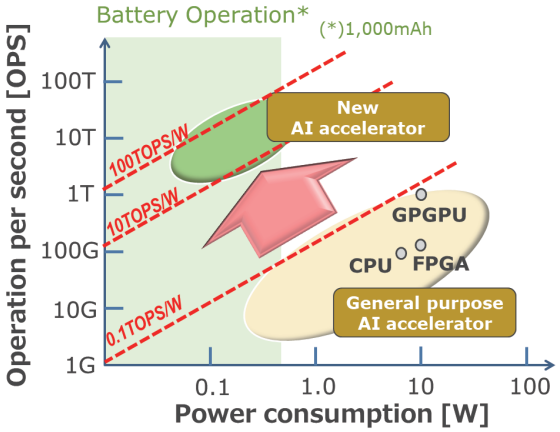

図 1　AI アクセラレータの特性（処理能力 vs 消費電力）

Figure 1. Characteristics of AI accelerators (Processing performance vs. Power consumption)

† 大阪国際工科専門職大学，大阪
International Professional University of Technology in Osaka, 3-3-1 Umeda, Kita-ku, Osaka 530-0001 Japan

a) E-mail: hayashikoshi.mas@iput.ac.jp

そのためには，演算軽量化のアルゴリズム，電力効率を最大限に活用できるアーキテクチャ，低電力メモリ素子デバイス技術の各階層間での技術連携が課題である．本論文では，電力効率を最大限に活用できる CiM アーキテクチャについて議論する．

2. 既存研究

表 1 にデジタル，および各種メモリ素子を用いた AI アクセラレータ [1-4] の特性比較を示す．メモリ素子を用いた CiM 技術を適用することで電力性能 (TOPS/W) の向上が期待される．CiM は，主に AI アクセラレータの分野で使用される技術の一つで，メモリ素子を使用して計算とデータの保存を同時に実行する．通常のデジタル計算機では，データをメモリから取り出し，プロセッサで計算し，再びメモリに書き戻すというプロセスに従うが，CiM ではメモリセル自体が計算を行うためデータの移動が不要となる．CiM はエッジ AI や種々のメモリでの実現が検討されている．

しかしながら，低電力化が期待できる次世代メモリ素子においてはデバイスのばらつきや特性変動等の課題があり，それを使いこなすためのアプローチが重要となる．そのためには，メモリ素子のデバイス特性を理解し，その過程で課題となるメモリ素子に適したエッジ向け組み込み用低電力アクセラレータにおけるニューラルネットワークアーキテクチャとその制御方法，エッジ向けとして実装可能な演算軽量化アルゴリズム等の技術の連携が重要となる．さらに，社会実装に向けては，軽量化アルゴリズムと低電力ニューラルネットワークアーキテクチャの組み合わせによる認識精度への影響を評価し，悪化する場合は認識精度補償の必要性の議論が必要となる．

3. 本研究のアプローチ

本研究では，AI アクセラレータに適したメモリ素子のデバイス特性に対する知見を踏まえた上で，低電力ニューラルネットワークアーキテクチャとその制御方法，演算軽量化するためのアルゴリズムを提案し，シミュレーションを用いて低電力化の効果と認識率への影響を模擬検証する．図 2 に低電力化のためのアプローチ例とそれに対応したハードウェア構成を示す．

入力画像に応じたスパース化の例として，入力差分検出と背景削除を想定し，変化に対応したアレイのみ活性化することで駆動エリアを最小化し低電力化を図る．スパース化とは，ニューラルネットワークの重みを疎行列とする最適化であり，入力画像を疎化することにより同様の効果が期待できる．

表 1　各種メモリ素子を用いた AI アクセラレータの特性比較
Table 1. Comparison of characteristics of AI accelerator using various memory devices

		Digital	CiM (Computing in Memory)			
		GPU (NVIDIA Xavier)	SRAM-CiM [1]	Flash-CiM [2]	MRAM-CiM [3]	ReRAM-CiM [4]
特徴 (アーキテクチャ)		演算器+memory	CiM	CiM	CiM	CiM
	重み	デジタル	デジタル(3値)	アナログ(8bits)	デジタル	アナログ(4bits)
	積和演算	デジタル(演算器)	アナログ(inメモリ)	アナログ(inメモリ)	アナログ(inメモリ)	アナログ(inメモリ)
演算制度		32bits	4~8bits	8bits	8bits	4bits
電力性能 (TOPS/W)		1	8.8	5.2	25.1	66.5
メモリ素子使いこなし技術	低電力化技術	Nearメモリ	CiMアーキ	CiMアーキ	CiMアーキ	CiMアーキ
	オンチップ推論補償技術	-	なし	なし	なし	なし
Availability	特性変動	なし	なし	あり	あり	あり
	プロセス習熟度	あり	あり	あり	開発中	開発中

[1] VLSI 2019 (Renesas), [2] ISSCC 2022 (Mythic), [3] ISSCC 2022 (NTHU), [4] VLSI 2018 (Panasonic)　Pros.　Cons.

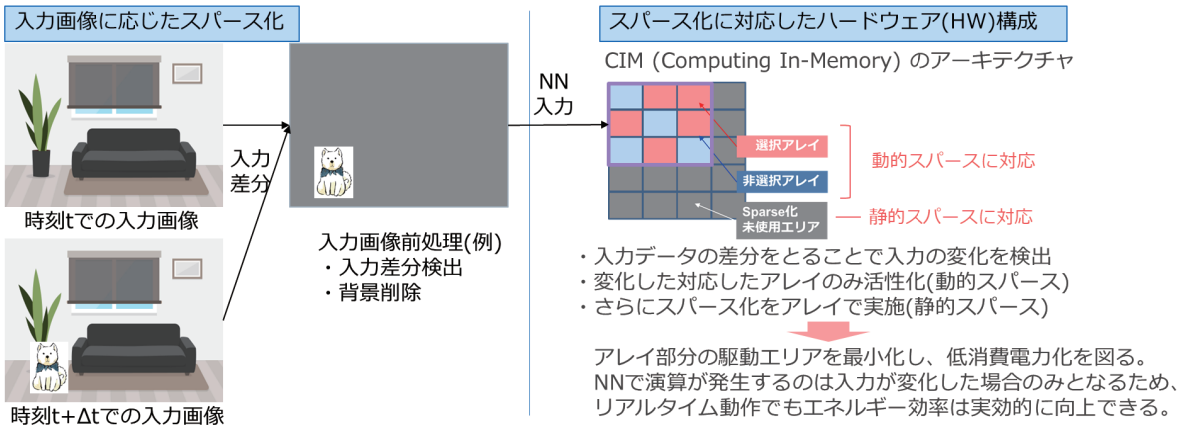

図 2　低電力化のためのアプローチとそれに対応したハードウェア構成
Figure 2. Approaches for lower power consumption and corresponding hardware configurations

4. 研究方法と結果

メモリ素子を用いたエッジ向け組み込み用低電力アクセラレータにおけるニューラルネットワークアーキテクチャとその制御方法，演算軽量化するためのアルゴリズムを検討し，電力削減効果と認識精度への影響を調査した．

4.1　ニューラルネットワーク構成

本研究では，図 3 に示す CNN2 層 +FCN3 層のニューラルネットワークを用いた．メモリ素子を用いた AI アクセラレータは AI シミュレーションで最適化したネットワークの動作を回路上で再現するよう設計されることから，AI シミュレーションの計算ログより消費電力の推定を行った．

4.2　シミュレーション結果

入力画像に応じたスパースの依存性を見るために，入力差分画像の占有比を振って電力削減効果を確認した．占有比は，入力画像に対する検出対象物の占める割合であり，検出対象物の画素をシュリンクすることで対応した．入力画像のデータセットは CIFER-10（380 枚 : 38 枚 x10 種類）を用いた．

結果を図 4 に示す．消費電力と推論精度は平均値（リファレンスの場合の結果で規格化）で集計した．

結果として，入力差分を検出し，背景削除すると

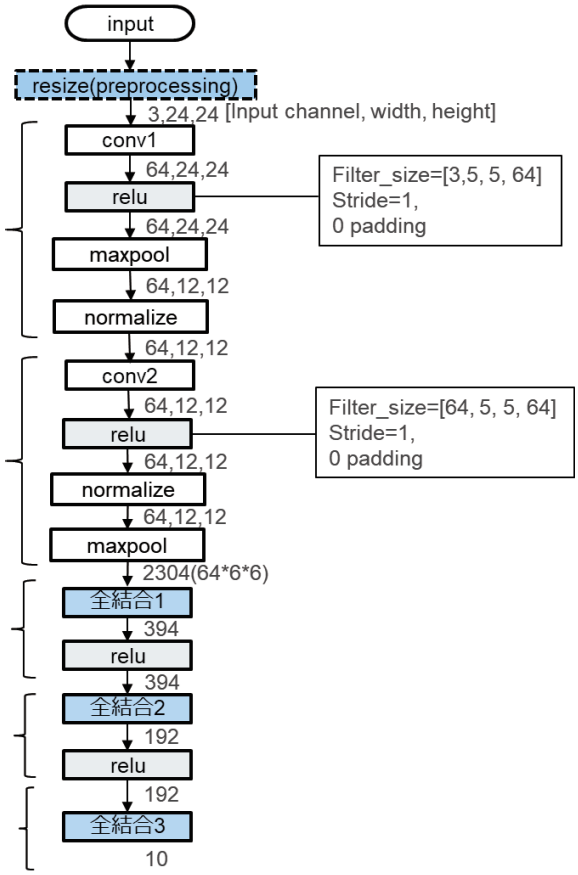

図 3　本研究で用いたニューラルネットワーク構成
Figure 3. Neural network configuration used in this research

ることで，占有比 100 ～ 14% に対して 22 ～ 72% の電力削減効果があることがわかり，入力画像に応じたスペースを行うことで低電力化が期待できる．ただし，推論精度は，占有比 100 ～ 14% に対して +14 ～ -83% で低下傾向にある．ここで，

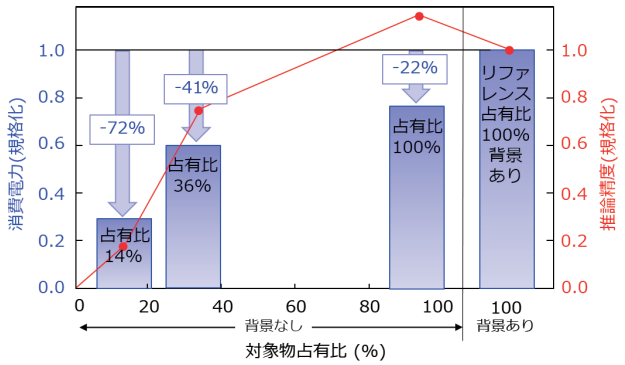

図 4　消費電力・推論精度の対象物占有比依存性

Figure 4. Dependence of power consumption and inference accuracy on object occupancy ratio

占有比 100% で推論精度が向上しているのは，入力画像に対して占有比 100% のままで背景削除していることに依るものと考えられる．社会実装に向けてはアプリケーションに応じた最適な占有率の選定や，推論精度補償手法を検討していく必要がある．

4.3　今後のアプローチ

NVM を用いた場合はメモリセルの信頼性に課題（データ保持特性の変動など）があり，電力性能と演算精度の両立はさらに難しくなる．NVM における信頼性の課題はデバイスの特性に依存するものであり，プロセス改善である程度緩和できるが完全に解決できる問題ではない．よって，メモリ素子を用いた CiM 技術を適用するためには，メモリ素子の特性を十分知り，それを使いこなすための技術を実装することが必須となる．今後，この知見を踏まえ，低電力 CiM アレイに対して推論補償技術を実装するための改良を行い，さらなる低電力性能の向上と認識精度の維持の両立を見込む．

5. むすび

本論文では，電力効率を最大限に活用できるエッジ AI 向け CiM アーキテクチャの実現可能性について議論してきた．本研究による次世代メモリ素子を用いたエッジ向け組み込み用低電力 AI アクセラレータを実現することで，従来のデジタル AI アクセラレータより大幅な電力削減が期待でき，Society5.0 以降必須となるエッジ・インテリジェンスの実現に大きく寄与できる点で社会的意義は大きいと考える．

謝　　辞

本研究は，前職であるルネサスエレクトロニクス株式会社での研究および JSPS 科研費 JP21K21298 の助成を受けて実施した内容をまとめたものである．ご支援いただき，深く感謝いたします．

文　　献

[1] S. Okumura, et al., “A Ternary Based Bit Scalable, 8.80 TOPS/W CNN accelerator with Many-core Processing-in-memory Architecture with 896K synapses/mm2,” IEEE Symposium on VLSI Circuits, 2019.

[2] L. Fick, et al., “Analog Matrix Processor for Edge AI Real-Time Video Analytics,” IEEE International Solid- State Circuits Conference, 2022.

[3] Y-C Chiu, et al., “A 22nm 4Mb STT-MRAM Data-Encrypted Near-Memory Computation Macro with a 192GB/s Read-and-Decryption Bandwidth and 25.1-55.1TOPS/W 8b MAC for AI Operations,” IEEE International Solid- State Circuits Conference, 2022.

[4] R. Mochida, et al., “A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture,” IEEE Symposium on VLSI Technology, 2018.

林越　正紀

1986 年神戸大学大学院 修士課程修了．同年三菱電機株式会社に入社．以降，同社 (現ルネサスエレクトロニクス株式会社) にて，半導体メモリ (DRAM,MRAM, 次世代 NVRAM) の開発・事業化に従事．2018 年金沢大学大学院 博士後期課程 修了 博士 (工学).

現在，大阪国際工科専門職大学にて，次世代メモリの AI(人工知能) 応用研究に従事．IEEE，IEEE SSCS，IEEE EDS 会員．