メモリエラーを考慮した Computation-in-Memory 向け ニューラルネットワーク精度評価シミュレータ

樋口和英 [1]　松井千尋 [1]　三澤奈央子 [1]　竹内健 [1]

概要：本研究では，包括的な CiM（Computation-in-Memory）シミュレータを提案する．このシミュレータは，CiM におけるメモリデバイスのセルの多値度，コンダクタンスのばらつきやシフトなどの様々なメモリデバイスの非理想性を模擬することができる．本研究では，VGG-16 と ResNet-34 のデバイス非理想性による推論精度の劣化を調査した．シミュレーションの結果，CiM の推論精度にはメモリデバイスにおけるコンダクタンスのランダムなばらつきよりも，コンダクタンスのシフトが重大な影響を与えることが判明した．

キーワード：Computation-in-Memory, Non-volatile Memory, Device Non-Ideality

1. はじめに

Computation-in-Memory（CiM）は，メモリアレイ構造を利用して乗算・累積（multiply-and-accumulate: MAC）演算を実現する．MAC 演算は，DNN の中で最も計算資源を消費する[1]. 図 1 (a)は，提案する精度評価シミュレータの概要を示したものである．このシミュレータでは，畳み込み層と全結合層における重みを任意に量子化し，その重みに任意の分布に従ったばらつきを付加することや，一定の値で加減させることができる．このように，DNN の重みを操作することで，CiM メモリセルにおけるデバイスの非理想性を再現することができる．

シミュレータにおいて，操作された重みの分布と推論の精度を得ることができる． 図 1 (b)は CiM の構造を示しており x，w，y はそれぞれ入力データ，重み，出力データを表している．これらの変数の分解能は，AD/DA コンバータの分解能や不揮発性メモリデバイスの MLC 動作によって制限される．CiM の重みは，図 1(c)に示すように，一様/非一様な変動やシフトなどの非理想性を持つコンダクタンスとして表される．

CiM 用の不揮発性メモリには, ReRAM, PRAM, MRAM, NAND フラッシュメモリ，FeFET などがある[2][3][4][5]. 各デバイスは，図 1 (c)に示されるように，異なる非理想性を持っている．コンダクタンスの一様/非一様なばらつきやシフトは，ベリファイ・プログラム，データ保持時エラー，リードディスターブなどが原因である．そこで，本研究では，実際のメモリデバイスの非理想性が DNN の推論精度に与える影響を調べるため，既存の CiM シミュレーションプラットフォーム[6][7]よりも，より重み分布を柔軟に操作できるシミュレータを提案する．

2. シミュレータによる重みの操作

本研究では, CIFAR-10 を ResNet-34 で学習したモデルと，MNIST を VGG-16 で学習したモデルを用いた[8]. 図 2 (a)は，学習した ResNet-34 の第 1 畳み込み層の重み分布を示している．本発表では，量子化，ばらつき，シフトを模擬する．図 2 (b)は図 2 (a)の重みに相当するメモリのコンダクタンスを MLC によって量子化した分布を示している．図 2 (c)は，図 2 (b)の分布に NAND フラッシュメモリのベリファイ・プログラムを想定し，正規分布に従うランダムなばらつきを加えた重みの分布を示している．図 2 (d)は，PRAM[3]のデータ保持エラーを想定して，図 2 (b)の分布を一様にシフトしたものである．

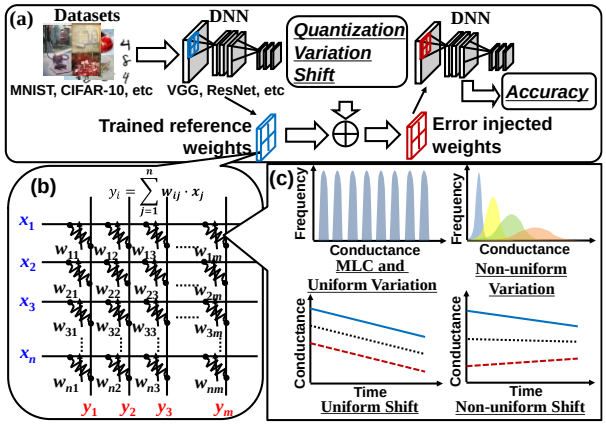

図 1 (a) 提案するシミュレーションプラットフォームの概要． (b) CiM の構造. (c) メモリデバイスの非理想性の例.

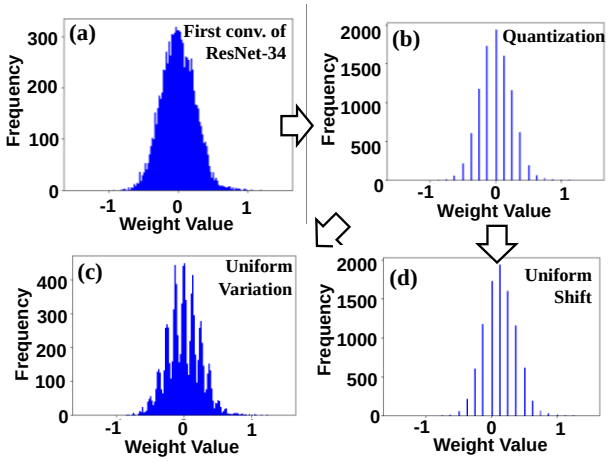

図 2　(a) ResNet-34 の第 1 層の畳み込み層における重み分布 (b)量子化した重み (c)正規分布に従うばらつきを付加した重み (d) 一定値を加算した重み

1 東京大学大学院
Graduate School of Engineering, The University of Tokyo

ⓒ2021 Information Processing Society of Japan

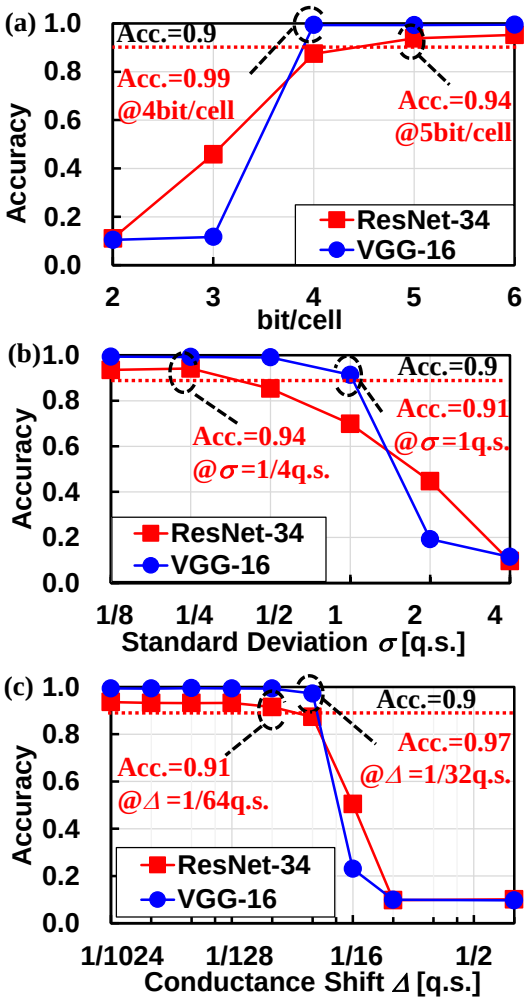

図 3(a) ビット/セルと推論精度 (b) 標準偏差σと推論精度 (c)正規分布に従うばらつきを付加した重み (d) 一定のシフト量を加算した重み

3. 非理想性による推論劣化の影響

以下の推論精度は，VGG-16 または ResNet-34 の全ての畳み込み層と全結合層の重みを操作した結果である．推論精度は 0.9 を目標としている．図 3 (a)は，１つの重みあたりの量子化粒度（ビット/セル）と推論精度の関係を示したものである． VGG-16 では 4 ビット/セル以上，ResNet-34 では 5 ビット/セル以上が必要であることが判明した．図 3 (b)は量子化された重みに正規分布に従う分散を加えた重みの標準偏差σと推論精度の関係を示したものである．VGG-16 では標準偏差σ を 1 量子化ステップ(q.s.)以下，ResNet-34 では 1/4q.s.以下に収めなければならないことが判明した．図 3 (c)は量子化された重みに加えたシフト量Δと推論精度の関係を示したものである．VGG-16 ではシフト量Δ を 1/32q.s.以下，ResNet-34 では 1/64q.s.以下にする必要があることが判明した．

以上から，実用的な推論を行うためには，メモリデバイスは，4 ビット/セル以上が必要であることがわかった．また，メモリデバイスにおけるコンダクタンスのばらつきよりもコンダクタンスがシフトしてしまう方が，推論精度が大きく劣化することが判明した．

4. おわりに

包括的な CiM シミュレータを提案した．このシミュレータでは，今回利用したデータセットの CIFAR-10 や MNIST，ニューラルネットワークモデルの VGG-16 や ResNet-34 に限らず，様々なデータセットと DNN を選択することができる．また，ばらつきやシフトメモリデバイスの非理想性をエミュレートすることを示した．0.9 の推論精度を達成するための MLC のビット/セル，コンダクタンスのばらつきに許容される標準偏差σ，コンダクタンスの許容されるシフト量Δを表 I にまとめた．シミュレーション結果から，実用的な推論には 4 ビット/セル以上が必要であること，コンダクタンスのばらつきよりもシフトの方が推論精度を大きく劣化させることがわかった．

謝辞 この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（ＮＥＤＯ）の委託業務の結果得られたものです．

表 1 シミュレーション結果のまとめ

	VGG-16 w/ BN on MNIST	ResNet-34 on CIFAR-10
Multi-Level Cell	≥ 4 bit/cell	≥ 5 bit/cell
Uniform Variation	σ ≤ 1 q.s.	σ ≤ 1/4 q.s.
Uniform Shift	Δ ≤ 1/32 q.s.	Δ ≤ 1/64 q.s.

参考文献

[1] S. Shukla et al., ”A Scalable Multi-TeraOPS Core for AI Training and Inference,” *L-SSC*, vol.1, no.12 pp. 217–220, 2018.
[2] R. Yasuhara et al., “Reliability Issues in Analog ReRAM Based Neural-Network Processor,” *IRPS*, 2019, pp. 1–5.
[3] Y. Lu et al., “Accelerated Local Training of CNNs by Optimized Direct Feedback Alignment Based on Stochasticity of 4 Mb C-doped Ge2Sb2Te5 PCM Chip in 40 nm Node,” *IEDM*, 2020, p. 36.3.1-36.3.4.
[4] K. Mizoguchi et al., “Data-Retention Characteristics Comparison of 2D and 3D TLC NAND Flash Memories,” *IMW*, 2017, pp 1-4.
[5] C. Matsui et al., “Application-Induced Cell Reliability Variability-Aware Approximate Computing in TaOx-based ReRAM Data Center Storage for Machine Learning,” *VLSI Tech.*, 2021.
[6] X. Peng et al., “DNN+NeuroSim: An End-to-End Benchmarking Framework for Compute-in-Memory Accelerators with Versatile Device Technologies,” *IEDM*, 2019, p. 32.5.1-32.5.4.
[7] L. Mei et al., “ZigZag: Enlarging Joint Architecture-Mapping Design Space Exploration for DNN Accelerators,” *TC*, vol.70, no.1, pp. 1160–1174, 2021.
[8] K. Higuchi et al., “Comprehensive Computation-in-Memory Simulation Platform with Non-volatile Memory Non-Ideality Consideration for Deep Learning Applications,” *SSDM*, pp. 121–122, 2021.

ⓒ2021 Information Processing Society of Japan