メモリデバイスの非理想性を考慮した 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

Neural Network Accuracy Evaluation Simulator for Computation-in-Memory with Non ideality of Memory Devices

東大工，　○樋口 和英[1]，松井 千尋[1]，三澤 奈央子[1]，竹内 健[1]

Univ. of Tokyo[1], ○Kazuhide Higuchi[1], Chihiro Matsui[1], Naoko Misawa[1], Ken Takeuchi[1]

E-mail: higuchi@co-design.t.u-tokyo.ac.jp

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）演算を行う．MAC 演算は，ディープニューラルネットワーク（DNN）の中で最も計算資源を消費する演算である[1]．Fig.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[2]．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． Fig.1(b)は CiM の構造を示しており，入力データ・重み・出力データの分解能は，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．各デバイスは，Fig.1(c)に示されるように，異なる非理想性を持っている．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，重み分布を Fig.2 に示すように柔軟に操作できるシミュレータを提案する．

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．

参考文献　[1] S. Shukla et al., *L-SSC*, vol. 1, no. 12, pp. 217-220, 2018. [2] K. Higuchi et al., *SSDM*, pp. 121-122, 2021.

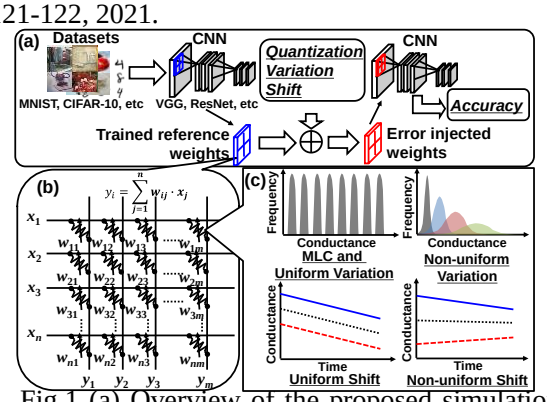

Fig.1 (a) Overview of the proposed simulation platform. (b) Structure of CiM. (c) Examples of memory device non-idealities [2].

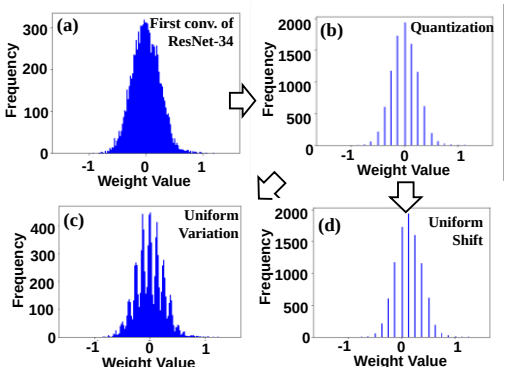

Fig.2 (a) Weight distribution of the first layer of ResNet-34. (b) Quantized weight distribution. (c) Adding Gaussian distribution to weight distribution. (d) Uniformly shifted weight distribution [2].